JEP 2006

Actes des XXVI^{es} Journées d'Etudes sur la Parole

Dinard, France

12-16 juin 2006

organisées par

l'Institut de Recherche en Informatique et Systèmes Aléatoires
(IRISA)

sous l'égide de

l'Association Francophone de la Communication Parlée
(AFCP)

branche régionale de

l'International Speech Communication Association
(ISCA)

Table des matières

Préface

Message du président de l'Association Francophone de la Communication Parlée

Chers collègues,

Après l'organisation de la conférence conjointe JEP-TALN en 2004 à Fès, la nouvelle édition des Journées d'Etude sur la Parole retrouve cette année son format plus traditionnel, recentré sur la communauté "Parole". Ce choix est celui des membres de l'Association Francophone de la Communication Parlée qui, dans leur majorité, ont souhaité privilégier l'alternance d'une conférence JEP-TALN, grand événement francophone traitant de tous les aspects du langage parlé et écrit, et d'une conférence indépendante consacrée plus spécifiquement au langage parlé.
Ce "resserrement" sur notre communauté, que vous avez souhaité, n'a donc très logiquement pas diminué votre intérêt pour les JEP, puisque nous atteignons sensiblement le même nombre de soumissions (environ 160) qu'il y a deux ans, parmi lesquelles environ 125 ont été retenues pour la conférence. Je me réjouis tout particulièrement de voir que le caractère international des JEP est maintenu pour cette édition, puisque, même si une forte majorité des contributions émane de laboratoires français, nous compterons parmi les participants des collègues venant d'Algérie, d'Allemagne, de Belgique, du Brésil, du Canada, du Maroc, de la Suisse, de la Tunisie et du Vietnam.
Outre son caractère francophone, les JEP présentent la particularité de ne pas avoir de sessions parallèles. C'est donc l'occasion quasi unique pour chacun d'entre nous d'assister à des exposés ou de voir des posters sur des domaines de la communication parlée qui ne sont pas au centre de nos travaux de recherche. Cette année le Comité de Programme est même allé encore plus loin dans cette direction, puisqu' aucune des sessions posters ne sera thématique et qu'on trouvera dans chacune d'entre elles un échantillonnage de chacun des grands domaines de notre recherche. Les JEP sont aussi traditionnellement l'occasion de faciliter les échanges entre les jeunes chercheurs et les chercheurs plus confirmés dans une ambiance simple et conviviale. La forte participation qui s'annonce (plus de 150 participants) et le choix judicieux du beau Manoir de la Vicomté à Dinard sur la Côte d'Emeraude par le comité d'organisation, permet de penser que toutes les conditions sont réunies pour que de ce point de vue l'édition 2006 des JEP soit aussi un succès,....

Avant de conclure, je souhaiterais remercier tout particulièrement Frédéric Bimbot, Olivier Boëffard, Marie-Noëlle Georgeault, Guillaume Gravier, Elisabeth Lebret et Laurent Miclet, pour le talent et l'enthousiasme qu'ils ont mis au service de l'organisation de cette conférence. Depuis le choix du lieu jusqu'à la réalisation des actes, ils ont fourni un travail énorme en interaction avec le Conseil d'Administration de l'AFCP, et ceci avec une efficacité, une disponibilité et une simplicité remarquables. Et ce n'est pas fini, puisque la conférence elle-même s'annonce riche en événements de diverses sortes,....

La relecture et la sélection des papiers a nécessité les contributions de très nombreux relecteurs, dans des délais plutôt courts. Tous ces relecteurs ont répondu dans les délais et dans un esprit très constructif. Qu'ils en soient aussi très sincèrement remerciés.

Bienvenue donc aux JEP2006. Qu'elles soient pour chacune et chacun d'entre vous l'occasion d'échanges scientifiques et personnels riches et fructueux.

Pascal Perrier,
Président de l'AFCP,
Président du Comité de Programme des JEP2006.

Message du président du comité d'organisation

C'est avec fierté et émotion que l'IRISA assure l'organisation de la 26ème édition des Journées d'Etudes sur la Parole, à Dinard, en cette mi-juin 2006. Fierté de voir converger en Bretagne un si grand nombre de collègues, issus des nombreuses disciplines des sciences de la Communication Parlée, basés en France ou dans d'autres pays francophones, nouveaux venus dans le domaine ou chercheurs confirmés.

Emotion d'être les artisans de ce rendez-vous qui rythme et structure la vie de notre communauté, permettant des échanges scientifiques en langue française, à intervalles réguliers, et des rencontres interdisciplinaires fructueuses et conviviales. D'ailleurs, qui ne se souvient pas de ses "premières" JEP ?

Ces 26èmes JEP ont été mises en place avec le souci de suivre au plus près les valeurs qui fondent cette conférence phare de l'Association Francophone de la Communication Parlée (AFCP).

A cet égard, un effort particulier du Comité d'Organisation a porté sur la cohésion de la conférence, par le choix d'un lieu unique pour les sessions scientifiques et l'hébergement, bâtiment de caractère à dimension humaine, bien situé mais un peu à l'écart du centre ville, propice aux rencontres et aux discussions sans s'abstraire totalement du contexte environnant.

L'accessibilité de la conférence a été la seconde préoccupation, avec des tarifs incitatifs, limitant la charge financière pour les laboratoires lointains ou fortement représentés. A cet égard, nous sommes reconnaissants aux nombreux sponsors qui, par leur aide financière, nous ont permis d'abaisser très notablement les frais d'inscriptions par rapport au prix de revient.

Enfin, la convivialité a été notre troisième objectif, en veillant à offrir aux participants des moments de détente et des événements sociaux nombreux et variés pour agrémenter les fins de journée et les soirées, après les échanges intenses prévus au programme scientifique.

A la date où j'écris ces lignes, le nombre élevé d'inscriptions enregistrées laisse supposer que cette édition des JEP sera l'une des plus importante par son volume. Le niveau d'exigence des relecteurs garantit pour sa part la qualité scientifique des communications.

En ce qui concerne le déroulement de la conférence, le professionnalisme des membres du service REV de l'IRISA (Elisabeth et Marie-Noëlle), l'énergie et la ténacité du principal artisan de ces JEP'2006 (Guillaume) et le concours solidaire de l'ensemble du Comité d'Organisation sont les meilleurs gages de succès de cette rencontre.

Nous aurons à cœur d'être à la hauteur de vos attentes.

Pour le Comité d'Organisation,

Le Président - Frédéric BIMBOT

Comité de programme

Laurent Besacier	CLIPS / GEOD, Grenoble
Frédéric Bimbot	CNRS - UMR 6074 IRISA, Rennes
Olivier Boëffard	Université de Rennes 1 - ENSSAT / IRISA, Lannion
Jean-François Bonastre	LIA, Avignon
Mohamed Embarki	Laboratoire de Phonétique, Montpellier
Jêrome Farinas	IRIT, Toulouse
Cecile Fougeron	Laboratoire de Phonétique et Phonologie, CNRS / Paris3
Alain Ghio	LPL, Aix-en-Provence
Hervé Glotin	SIS, Toulon
Guillaume Gravier	CNRS - UMR 6074 IRISA, Rennes
Irina Illina	LORIA, Nancy
Lori Lamel	LIMSI, Orsay
Fabrice Lefevre	LIA, Avignon
Georges Linarès	LIA, Avignon
Sylvain Meigner	LIUM, Le Mans
Laurent Miclet	Université Rennes 1 - ENSSAT / IRISA, Lannion
François Pellegrino	UMR5596 Dynamique Du Langage, CNRS - Université Lyon 2
Pascal Perrier	ICP, Grenoble (Président)
Solange Rossato	ICP, Grenoble
Jean Schoentgen	ULB, Bruxelles
Rudolph Sock	IPS, Strasbourg

Comité d'organisation

Frédéric Bimbot (Président)
Olivier Boëffard
Marie-Noëlle Georgeault
Guillaume Gravier
Elisabeth Lebret
Laurent Miclet

Remerciements

Les membres du comité de programme tiennent à remercier les personnes suivantes pour l'aide précieuse qu'ils ont apportés lors de la relecture des nombreux articles soumis aux Journées d'Étude sur la Parole. Par ordre alphabétique :

Christian Abry (ICP, Grenoble)
Gilles Adda (LIMSI, Orsay)
Christophe d'Alessandro (LIMSI, Orsay)
Alexandre Allauzen (LIMSI, Orsay)
Véronique Aubergé (ICP, Grenoble)
Pierre Badin (ICP, Grenoble)
Melissa Barkat-Defradas (ICAR-Praxiling)
Claude Barras (LIMSI, Orsay)
Frédéric Bechet (LIA, Avignon)
Nidhal Ben Aloui (LSIS, Toulon)
Frédéric Berthommier (ICP, Grenoble)
Anne Bonneau (LORIA, Nancy)
Philippe Boula de Mareuil (LIMSI, Orsay)
Christophe Cerisara (LORIA, Nancy)
Vincent Colotte (LORIA, Nancy)
Lise Crevier-Buchman (LPP & HEGP)
Elisabeth Delais-Roussarie (LLF)
Paul Deléglise (LIUM, Le Mans)
Véronique Delvaux (Laboratoire de Phonologie, ULB)
Albert Di Cristo (LPL, Aix-en-Provence)
Robert Espesser (LPL, Aix-en-Provence)
Yannick Estève (LIUM, Le Mans)
Dominique Fohr (LORIA, Nancy)
Corinne Fredouille (LIA, Avignon)
Jean-Luc Gauvain (LIMSI, Orsay)
Cédric Gendrot (LPP)
Laurent Girin (ICP, Grenoble)
Pierre Hallé (LPP)
Jean-Paul Haton (LORIA, Nancy)
Nathalie Henrich (ICP, Grenoble)
Bruno Jacob (LIUM, Le Mans)
Barbara Kühnert (LPP & Inst. du Monde Anglophone)
David Langlois (LORIA, Nancy)
Yves Laprie (LORIA, Nancy)
Bertrand Lauret (Paris 3 - Sorbonne Nouvelle)
Jean-Sylvain Liénard (LIMSI, Orsay)
Hélène Loevenbruck (ICP, Grenoble)
Daniel Luzzati (LIUM, Le Mans)
Shinji Maeda (ENST)
Driss Matrouf (LIA, Avignon)
Christine Meunier (LPL, Aix-en-Provence)
Fanny Meunier (DDL, Lyon)
Yohann Meynadier (LPL, Aix-en-Provence)
Irina Nesterenko (LPL, Aix-en-Provence)
Noël Nguyen (LPL, Aix-en-Provence)
Yuki Nishinuma (LPL, Aix-en-Provence)
Pascal Nocera (LIA, Avignon)
Michel Pitermann (LPL, Aix-en-Provence)
Crystel Portes (LPL, Aix-en-Provence)
Béatrice Priego (LPL, Aix-en-Provence)
Annie Rialland (LPP)
Rachid Ridouane (LPP)
Albert Rilliard (ICP, Grenoble)
Christophe Savariaux (ICP, Grenoble)
Jean-Luc Schwartz (ICP, Grenoble)
Bernard Teston (LPL, Aix-en-Provence)
Jacqueline Vaissière (LPP)
Nathalie Vallée (ICP, Grenoble)
Annemie Van Hirtum (ICP, Grenoble)
Béatrice Vaxelaire (IPS, Strasbourg)
Anne Vilain (ICP, Grenoble)
Tuan Vu-Ngoc (LIMSI, Orsay)
Pauline Welby (ICP, Grenoble)
François Yvon (Télécom Paris)

Session I

Introduction

Lundi 12 juin 2006 - 11h30 12h30

Phonétique et Phonologie au siècle des Lumières

Christophe Rey

Equipe DELIC, Université de Provence
Christophe.Rey@up.univ-aix.fr
http://www.up.univ-mrs.fr/delic/perso/rey/index.html

ABSTRACT

The present paper proposes a presentation of various aspects of Nicolas Beauzée's theories on French sounds. Remained unexploited, Beauzée's theories sum up the best knowledge available in XVIII th century, before the progress brought by Comparatism and Dialectology, and found the description of the sounds of the language as a true field of study for grammatical science. We will also show that these same theories represent the first steps of a properly phonological reflexion.

1. INTRODUCTION

Historiquement, la Phonétique ne semble s'être constituée comme discipline scientifique véritable qu'au XIX[e] siècle, en bénéficiant notamment des travaux du Comparatisme et de l'essor de la Dialectologie. En tant qu'historien de la langue et plus précisément de la Grammaire, nous nous sommes intéressé [7] à la question de la nature des descriptions phoniques de la langue française avant l'émergence de cette discipline. En prenant pour cadre d'étude les théories développées par les grammairiens-philosophes de l'*Encyclopédie* ou *dictionnaire raisonné* (1751-1772) [6] et de l'*Encyclopédie Méthodique* (1782-1832) [2], nous avons voulu - ainsi que l'a fait Geneviève Clérico [4] pour le XVI[e] siècle - dresser un panorama des connaissances relatives à la description des sons de notre langue au siècle des Lumières. L'un des buts de cette expertise était de déterminer le rôle joué par les théories du XVIII[e] siècle dans l'établissement de notre Phonétique moderne.

A défaut d'avoir découvert chez les grammairiens-philosophes français les fondateurs avérés de cette discipline, nous avons pu mettre en évidence les théories novatrices du grammairien Nicolas Beauzée. Héritières des avancées précédentes, ces dernières synthétisent non seulement les connaissances les plus abouties du siècle, mais apportent également un plus grande technicité et scientificité dans la description phonique de la langue.

Après une présentation de certaines avancées apportées par Beauzée dans la description articulatoire des sons, nous souhaitons évoquer ici les développements théoriques qui chez ce dernier constituent, selon nous, les prémices véritables de notre Phonologie moderne, ou tout au moins ce que Sylvain Auroux [1] appelle une épiphonologie.

2. LA DESCRIPTION ARTICULATOIRE

Décrire les sons de la langue française au XVIII[e] siècle, que cela soit sous la plume de Nicolas Beauzée ou de ses contemporains, revient à dresser et à étudier l'inventaire des sons que transcrit notre orthographe. Dans cette tentative de délimitation des unités phoniques de notre langue, Beauzée, notamment à travers les théories qu'il développe dans sa *Grammaire générale* (1767) [3], apparaît comme nous allons nous attacher à le démontrer ci-dessous, comme le grammairien fournissant la description la plus technique et la plus proche de la nôtre.

2.1. Une classification véritable

Alors que chez ses prédécesseurs et contemporains la classification des sons de la langue est fournie de manière linéaire et relativement peu structurée, Beauzée fournit une description qui constitue - pour la langue française en tout cas - la première de ce type. Il s'agit d'une organisation sous forme de schémas que nous avons reproduits ci-dessous. La figure 1 concerne ce que Beauzée appelle les *voix*, c'est-à-dire les unités vocaliques, et la figure 2 concerne les *articulations*, autrement dit les unités consonantiques.

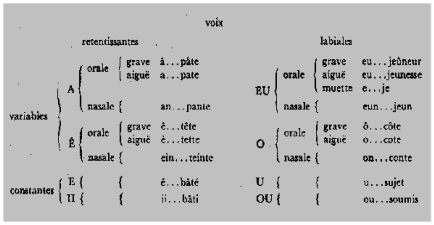


Figure 1 : Le système des voix dans la *Grammaire Générale* de Beauzée

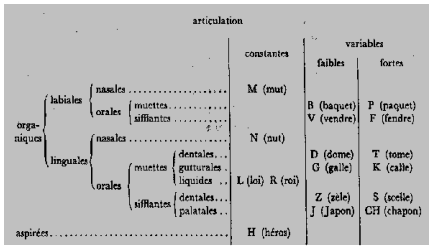


Figure 2 : Le système des articulations dans la *Grammaire Générale* de Beauzée.

Ne pouvant nous attarder sur ces deux schématisations, nous pouvons néanmoins mentionner leur aspect arborescent et finement structuré, reposant à la fois sur des oppositions en fonction de lieux et de modes d'articulation. Il s'agit là d'une méthode de classification instaurant une rupture évidente avec les travaux précédents et érigeant l'étude des aspects phoniques de la langue comme un champ d'analyse véritable de la science grammaticale.

2.2. Des apports fondamentaux

Intéressons-nous à présent aux avancées significatives apportées par Beauzée dans la description des sons du français qu'il propose. Pour cela, considérons au moins deux caractéristiques distinctives que ce dernier paraît introduire et que notre système actuel semble avoir conservé : l'opposition *orales/nasales* et l'opposition *muettes/sifflantes*.

L'opposition orales/nasales

Rendue possible grâce à la découverte des voyelles nasales par l'abbé Louis-Courcillon de Dangeau au XVII^e siècle [5], l'opposition entre des unités phoniques *orales* et des unités phoniques *nasales* ne s'est pas imposée tout de suite chez les grammairiens du siècle des Lumières. Elle ne se traduit en effet avant Beauzée, que sous la forme d'une séparation entre les unités *nasales* et les autres unités. L'auteur de la *Grammaire générale* est le premier à fournir une opposition entre des unités *nasales* et des unités *orales*, formulant ainsi pour la première fois une thématisation lexicale de cette opposition.

"I°. Les articulations nasales sont celles qui font réfluer par le nez une partie de l'air sonore dans l'instant de l'interception, de manière qu'au moment de l'explosion il n'en reste qu'une partie pour produire la voix articulée. [...]

2°. Les articulations orales sont celles qui ne contraignent point l'air sonore de passer par le nez dans l'instant de l'interception, de manière qu'au moment de l'explosion tout sort par l'ouverture ordinaire de la bouche." [3]

L'opposition muettes/sifflantes

Au sein de sa classification des sons, Beauzée propose un dégroupement particulièrement intéressant qui lui non plus n'est pas attesté chez ses contemporains. Il s'agit de l'opposition *muettes/sifflantes*.

Reprise des grammairiens anciens, cette opposition n'est pas formulée chez les contemporains de Beauzée, lesquels se contentent de mentionner l'existence d'unités *muettes* sans les opposer à un autre type d'unités. Beauzée, pour sa part, fait de cette opposition l'un des piliers de son système et explicite même les caractéristiques articulatoires de ses unités *muettes* et *sifflantes* :

"Les articulations orales muettes sont celles qui naissent d'une interception totale de l'air sonore; de manière que, si la partie organique qui est mise en mouvement restoit dans l'état où ce mouvement la met d'abord, il ne pourroit s'échapper aucune partie de l'air sonore & l'on ne pourroit rien faire entendre de distinct." [3]

"Les articulations orales sifflantes sont celles qui naissent d'une interception imparfaite de l'air sonore; de manière que, quand la partie organique qui est mise en mouvement resteroit dans l'état où ce mouvement la met d'abord, il s'échapperoit pourtant assez d'air sonore pour faire entendre l'articulation même dont il s'agit, et même pour la faire durer longtemps comme une sorte de sifflement, de même que l'on fait durer les voix simples aussi longtemps que les poumons peuvent fournir de l'air [...]"[3]

Les propriétés articulatoires énoncées par Beauzée laissent clairement penser que son opposition entre articulations *muettes* et articulations *sifflantes* correspond à notre opposition moderne entre consonnes *occlusives* et consonnes *fricatives*. La répartition des articulations proposées dans la figure 2 corrobore d'ailleurs cette intuition.

Associée à l'opposition *orales/nasales*, cette seconde opposition contribue à faire du système de Beauzée non seulement la théorisation la plus séduisante et la plus aboutie du siècle des Lumières, mais aussi un système assez proche du nôtre.

3. REFLEXIONS PHONOLOGIQUES

Au delà des avancées significatives qu'il semble avoir apportées en termes de description articulatoire et de classification des sons, Beauzée apparaît également comme un personnage clé pour la diffusion de réflexions s'apparentant à des considérations d'ordre phonologiques.

3.1. L'opposition fortes/foibles

Au sein de sa classification des sons, Beauzée formule une opposition entre les articulations *foibles* et les articulations *fortes*. Cette opposition se trouve pour la première fois formulée par l'abbé de Dangeau :

"BP, VF, DT, GK, ZS, JCh

La première colonne est des lettres qu'on peut nommer foibles, & l'autre de celles qu'on peut nommer fortes : la première est de celles qui sont précédées par une petite émission de voix, & l'autre est de celles qui n'en ont point." [5]

Bien que la notion de vibration des cordes vocales ne soit à aucun moment évoquée dans les processus de production et de distinction des sons que Dangeau identifie, la répartition des sons proposée par ce dernier s'avère tout à fait troublante puisqu'elle correspond

exactement au dégroupement que nous opérons aujourd'hui entre nos sons voisés et non-voisés. En évoquant la présence d'une petite *émission de voix* au début de la production des consonnes *foibles* - émission qui pourrait avoir été suggérée par l'impression laissée lors de l'adduction des cordes vocales - Dangeau semble avoir pris conscience d'une différence articulatoire fondamentale entre les consonnes *foibles* et les consonnes *fortes*.

Dans sa *Grammaire Générale*, Beauzée apporte pour sa part une définition de l'opposition *foibles/fortes* qui n'est pas calquée sur celle de Dangeau mais qui nous laisse également croire qu'il pourrait s'agir de notre opposition entre sons *voisés* et *non-voisés*. Sa définition s'appuie sur l'idée d'une différence de "force" dans la production des sons :

"(N.) FOIBLE, adj. [...] On appelle foibles celles qui n'interceptent pas la voix avec toute la vigueur dont est capable la résistance de la partie organique qui en est le principe. B, V, D, G, Z, J, sont des articulations variables foibles. Voyez ARTICULATION & FORT." [2]

"(N.) FORT, E, adj. [...] On appelle fortes, celles qui interceptent la voix avec toute la vigueur dont est capable la résistance de la partie organique qui en est le principe. P, F, T, K, S, CH, sont des articulations fortes. Voyez ARTICULATION & FOIBLE." [2]

Tout comme chez Dangeau, cette opposition pourrait s'apparenter à notre opposition entre sons voisés et non-voisés ; nous ne nous attarderons pas sur ce point soumis à controverse [1] mais souhaitons plutôt insister sur le fait que cette opposition articulatoire va être illustrée par l'élaboration d'un certain nombre d'exemples constituant ce que nous appelons aujourd'hui des paires minimales.

Cette opposition se trouve d'abord illustrée au sein de l'article CONSONNE de l'*Encyclopédie* de Diderot et d'Alembert, rédigé par le grammairien César-Chesneau Dumarsais.

Table 1 : Illustration de l'opposition articulations faibles et articulations fortes chez Dumarsais

Consonnes foibles	**Consonnes fortes**
B	**P**
Beau. Bécher.	Peau. Pécher.
D	**T**
Dard. Doge.	Tard. Toge.
G, gue.	**C dur, K, ou Q, que.**
Gache. Gage.	Cache. Cage.
J, je.	**Ch, che.**
Jarretière.	Charretière.
Jatte.	Chatte.
V, ve.	**F, fe.**
Vain.	Faim.
Z, ze.	**S, se.**
Zèle.	Selle.

A la suite de Dumarsais, Beauzée illustre et explicite plus précisément encore cette opposition.

Constituant, comme nous le verrons plus bas, l'une des distinctions essentielles retenues dans sa classification des sons consonantiques, cette opposition est décrite et mise en exergue à plusieurs reprises dans sa *Grammaire générale* et dans ses articles de l'*Encyclopédie Méthodique* :

"GUTTURAL [...] Les articulations gutturales sont celles qui font retentir l'explosion de la voix dans la région du gosier. Il y en a deux bien sensibles dans le françois, G & Q ; telles qu'on les entend dans les mots **Gale**, **Cale** ; **vaguer**, **vaquer** ; &c." [2]

"PALATAL [...] Les articulations palatales sont des articulations linguales sifflantes, dont le sifflement s'exécute dans l'intérieur de la bouche, entre le milieu de la langue & le palais. Il y en a deux en francois, j & ch, telles qu'on les entend au commencement des mots **Japon**, **chapon**. " [2]

Les exemples reproduits ci-dessus montrent clairement que le grammairien dépasse la simple opposition phonétique des unités comparées pour envisager leur rôle d'unités pertinentes pouvant générer des oppositions de sens dans la langue.

3.2. L'utilisation généralisée des paires minimales

Au delà de l'opposition entre articulations *foibles* et articulations *fortes* amenant un certain nombre de réflexions de nature phonologique, les théories de Beauzée sont caractérisées par une utilisation conséquente et surtout scientifique de l'illustration sous forme de paires minimales.

En effet, à l'intérieur même de la classification schématique des sons qu'il propose, Beauzée a systématiquement recours à l'illustration des unités qu'il identifie sous une forme lexicale. Chaque son est mis en contexte au sein d'un lexème, permettant ainsi au lecteur de disposer d'une représentation phonique de l'élément traité.

En ce qui concerne son système vocalique, Beauzée oppose successivement [a], [ɑ], et [ã] présents dans les mots *pâte*, *pate*, et *pante*, [ɛ], [ɛ], et [ẽ], présents dans les mots *tête*, *tète*, et *teinte*, [œ], [ø], [ə] et [œ̃], figurant dans les mots *jeûneur*, *jeunesse*, *âge*, *jeun*, et enfin [o], [ɔ], et [õ], rencontrés dans les mots *côte*, *cote*, et *conte*.

Ces exemples montrent que le recours aux paires minimales caractérise les sons dont les propriétés

articulatoires sont les plus proches. Ce recours ne s'arrête toutefois pas à ces seules unités puisqu'il sert également à illustrer l'opposition entre [e] et [i] rencontrée dans les mots *bâté* et *bâti* et anciennement illustrée dans l'*Encyclopédie* par la paire *présent/prison*. Il est regrettable que ce changement particulièrement significatif opéré dans la *Grammaire générale* n'ait pas également bénéficié à l'opposition entre les sons [y] et [u], opposition toujours illustrée par la paire *sujet/soumis*.

Dans la classification des sons consonantiques de Beauzée, le recours aux paires minimales repose sur la distinction de deux catégories de sons : les articulations *variables* et les articulations *constantes*, c'est-à-dire entre les articulations susceptibles d'être *foibles* ou *fortes* et toutes les autres.

A aucun moment mises en parallèle dans la classification, les articulations M et N puis L et R sont respectivement illustrées par la paire *Mut/Nut* - remplacée plus tard par la paire *Mort/Nord* - et par la paire *Loi/Roi*. Beauzée emploie donc clairement une exemplification visant à souligner la valeur phonologique de ces différentes unités.

L'opposition entre articulations *foibles* et articulations *fortes* donne pour sa part lieu à des exemplifications telles que celles rencontrées dans l'article CONSONNE de l'*Encyclopédie*. Le tableau ci-dessous récapitule les différentes oppositions lexicales retenues.

Table 2 : Illustration de l'opposition articulations faibles et articulations fortes chez Beauzée.

Articulations variables	
articulations faibles	**articulations fortes**
B (baquet) **V** (vendre)	**P** (paquet) **F** (fendre)
D (dome) **G** (galle)	**T** (tome) **K** (calle)
Z (zèle) **J** (Japon)	**S** (scelle) **CH** (chapon)

Ce tableau achève d'illustrer le fait que le recours aux paires minimales est un procédé résolument ancré dans les théories de Beauzée. Il ne s'agit certes pas d'une procédé scientifique propre à ce grammairien, mais il s'agit d'un procédé non seulement particulièrement généralisé chez lui, mais dont l'un des mérites essentiels est d'être utilisé à un endroit aussi stratégique que la classification des sons de la langue identifiés.

4. CONCLUSION

A défaut d'avoir révolutionné les connaissances sur les sons de la langue française, Nicolas Beauzée s'impose comme l'un des grammairiens qui, avant l'émergence du Comparatisme et de la Dialectologie, a le mieux su exploiter les connaissances de ses prédécesseurs et de ses contemporains pour faire de l'étude des sons de la langue un domaine d'étude à part entière de la Grammaire.

La mise en avant, pour la première fois, d'un système entièrement structuré de l'aspect phonique de notre langue, marque l'aboutissement d'une maturation issue de plusieurs siècles d'étude. Héritier des Meigret, Ramus, Dangeau, Duclos, Dumarsais et autres, l'auteur de la *Grammaire Générale*, notamment à travers sa thématisation lexicale de l'opposition *oralité/nasalité*, mais aussi à travers sa prise en considération de l'opposition *muettes/sifflantes*, apporte une dimension technique jamais atteinte à l'étude des sons du français.

Bien que n'ayant pas eu d'influence explicite sur le développement de notre phonétique moderne, les travaux de Beauzée, grâce entre autres au recours généralisé et stratégique que fait ce dernier des paires minimales, nous ont fait basculer d'une morphophonologie assez répandue au XVIII[e] siècle, vers une épiphonologie préfigurant la phonologie à venir.

5. BIBLIOGRAPHIE

[1] S. Auroux. Note sur les progrès de la phonétique au XVIII[e] siècle, In *Histoire des idées linguistiques*, Tome 2, Philosophie et Langage, Mardaga, pp. 598-606, 1992.

[2] N. Beauzée, J-F. Marmontel. *Encyclopédie Méthodique. Grammaire & Littérature*. A Paris (chez Panckoucke), Liège (chez Plomteux). 3 vol, 1782-1784-1786.

[3] N. Beauzée. *Grammaire générale ou exposition raisonnée des éléments nécessaires du langage, pour servir de fondement à l'étude de toutes les langues*, Paris : J. Barbou, réédité en fac-similé, Stuttgard-Bad Cannstatt : Friedrich Fromann Verlag, 1974 (1767).

[4] G. Clérico. *Analyses phoniques et prosodiques au XVI[ème] siècle. Origine et préhistoire d'une discipline.* Thèse de doctorat d'Etat, Université de Paris VIII Saint-Denis, 1995.

[5] L-C. Dangeau. *Essais de grammaire* (1694), repris dans *Opuscules sur la langue françoise* (1754).

[6] D. Diderot, J Le Rond d'. Alembert. *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de Lettres*, Stuttgart, F. Frommann Verlag – G. Holzboog, 1990 (1751-1766).

[7] C. Rey. *Analyse et informatisation des articles traitant de l'étude des sons dans le dictionnaire Grammaire & Littérature de Nicolas Beauzée et Jean-François Marmontel, issu de l'Encyclopédie Méthodique*. Thèse de doctorat. Aix-en-Provence, 2004.

A la poursuite de la trace du signal de parole

Bernard Teston

Laboratoire Parole et Langage, UMR 6057 CNRS, Université de Provence, Aix-en-Provence
teston@lpl.univ-aix.fr

ABSTRACT

At the beginning of the nineteenth century, the linguists, physiologists and acoustics experts had only one goal: to make the speech visible to be able to study its nature and its structure. Many scientists and inventors then will often launch out to the continuation of the speech signal with very varied but not very effective techniques, during a half of the century. However, these sometimes curious devices will allow the researchers today to make fundamental discoveries on which our speech domain is founded.

1. INTRODUCTION

Au début du 19ème siècle, les linguistes étudient l'évolution historique des langues isolément sur des textes, les lettres de leurs alphabets, et non sur les sons qu'elles représentent. Une rupture épistémologique se manifeste alors par l'émergence, à la suite de la redécouverte des textes védiques, d'un mouvement qui pousse certains linguistes à porter leur attention sur le langage tel qu'il est dit plutôt qu'écrit. Cela nécessite, pour compléter les données de nos sens, des approches méthodologiques nouvelles empruntées à la physiologie (étude des articulations) et à la physique (acoustique des sons). Les savants de cette époque ne peuvent cependant que constater que les sons, tout comme les mouvements, sont des phénomènes physiques si fugaces que la connaissance de leurs mécanismes ne pourra se développer qu'à la condition d'inventer des techniques nouvelles pour les capturer et les restituer, *Verba volent scripta manent*. Ils vont donc être nombreux à travailler sur l'inscription graphique et l'enregistrement de la parole pour la rendre visible. Malgré toute l'imagination, l'énergie et la passion mises en œuvre par les différents protagonistes de cette aventure scientifique, les progrès seront très lents et se développeront sur près d'un siècle, à travers de multiples impasses techniques autant que de nombreuses querelles scientifiques ou d'intérêts financiers. Le qualificatif d'aventure scientifique convient bien à cette quête, à cette course à la trace, qu'est la poursuite du signal de parole qui alla même pour certains jusqu'à l'obsession. Mais ces progrès laborieux vont permettre cependant aux physiciens physiologistes et linguistes associés à cette aventure, dont de nombreux français, de fonder sur des bases solides nos connaissances actuelles et ceci au moyen d'une technologie exclusivement mécanique. C'est pour rendre hommage à ces facteurs d'instruments, ancêtres de nos microphones modernes, éditeurs de signaux et enregistreurs audio-numériques que nous présentons cette étude.

En 1807, le savant anglais Thomas Young, physicien, médecin et linguiste polyglotte, inscrit les vibrations d'un diapason sur la surface d'un cylindre tournant enduit de noir de fumée. C'est le premier enregistrement attesté d'une manifestation sonore. On doit à Young, qui avait toutes les compétences pour mener des études sur les mécanismes de la parole, des travaux importants dont une théorie sur la production des voyelles. Pour cela, il va tenter d'appliquer ce qui sera appelé plus tard la *méthode graphique* à l'inscription du signal de parole. On sait qu'il fit plusieurs tentatives dans ce sens sans jamais aboutir, bien qu'il ne les ait jamais mentionnées.

2. LE PHONAUTOGRAPHE

On ignore par quel dispositif Young remplaça le diapason pour tenter d'inscrire le signal de parole sur le cylindre mais c'est sur ce point que fut réalisé le saut technologique suivant, cinquante ans plus tard, par Léon Scott de Martinville. Ce dernier, typographe de métier à Paris, était passionné par l'*impression des phénomènes sonores* et particulièrement de la parole. Il avait dans l'idée de sténographier le discours d'un orateur. Pour ce faire, il imagina un dispositif inscripteur constitué par un pavillon et une membrane souple, inspiré du modèle anthropomorphique de l'oreille.

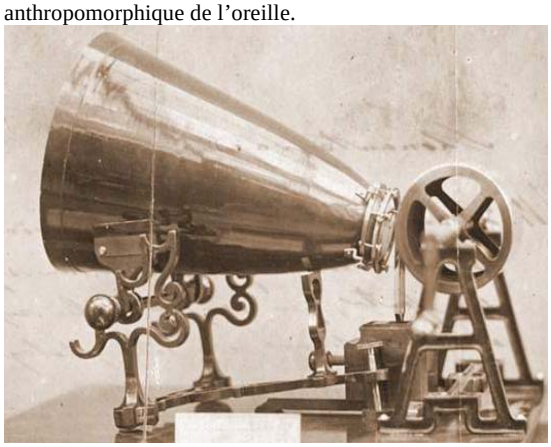


Figure 1 : Phonautographe de Scott-Koenig (Collection particulière).

Le pavillon amplifiait la pression acoustique des sons prononcés devant son ouverture, qui faisaient vibrer la membrane, dont les mouvements étaient gravés par l'intermédiaire d'un stylet, sur un cylindre tournant enduit de noir de fumée identique à celui de Young. Avec cet instrument, le *Phonautographe*, Scott obtint en 1858 les premières traces d'un signal de parole que l'on pouvait voir et conserver. Mais les résultats furent jugés très médiocres et entachés par de multiples modes résonants du pavillon et de l'ensemble membrane-stylet. Pour l'améliorer, il s'associera avec Rudolph Koenig qui en réalisera les versions les plus performantes jusqu'en 1875.

Le *Phonautographe* a été ainsi, et malgré ses imperfections, le premier enregistreur graphique capable de fournir une trace du signal de parole mais qui ne permettait pas sa restitution. Bien que Scott n'ait publié aucune expérience scientifique exécutée au moyen de son appareil, ce dernier a été abondamment utilisé par de nombreux chercheurs sur la parole tels que Frans Donders, Graham Bell, Hermann Helmholtz et Thomas Edison. L'appellation de *Phonautographe* est même devenue générique pour nommer tous les enregistreurs graphiques des sons ayant succédé à l'original de Scott et basé sur le principe de la membrane vibrante associée au stylet inscripteur.

Figure 2 : Membrane, aiguille, tambour et trace d'un signal du *Phonautographe* de Scott (collection particulière).

A la suite de Scott, de nombreux chercheurs apportèrent des améliorations multiples au *Phonautographe*.En 1872, l'Ecossais Graham Bell grava le signal sur une plaque de verre enduite de noir de fumée et soumise à un déplacement linéaire. Cela améliora de manière très significative la qualité de la trace. Des perfectionnements des membranes et des stylets graveurs permirent également d'obtenir des traces de signaux de parole plus détaillées et plus fidèles (figure 3). Ainsi, le Suisse Heinrich Schneebeli réalisa en 1878 les tracés, considérés comme faisant partie des meilleurs, obtenus au moyen de l'association membrane-stylet graveur. Grâce à leur finesse, il put leur appliquer le théorème de Fourier et en faire, pour la première fois, l'analyse harmonique.

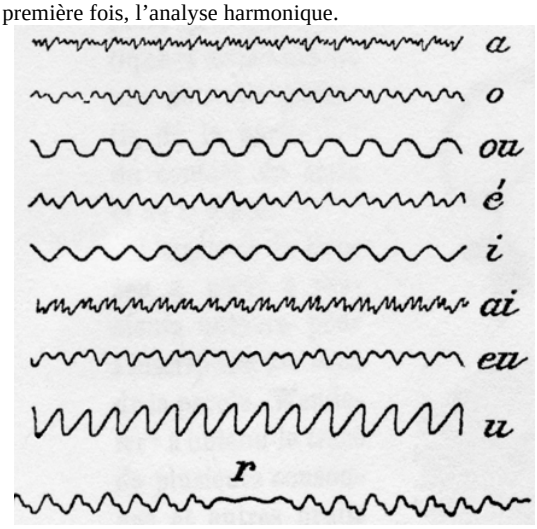


Figure 3 : Traces du signal de différentes voyelles du Phonautographe de Schneebeli en 1878, (d'après Marey, 1898).

L'ultime perfectionnement du principe du *Phonotographe* fut réalisé par l'américain Whitney Blake en 1878, avec la suppression du système de gravure constitué par le tandem stylet-papier noirci, source de multiples résonances et distorsions mécaniques. Il le remplaça par un système d'inscription optique constitué d'un petit miroir actionné par la membrane qui réfléchissait un rayon lumineux sur une plaque photographique. Ce fut le second saut technologique de cette saga.

3. L'AVENTURE DU PHONOGRAPHE

En 1877, le français Charles Cros présente à l'Académie des Sciences, et dépose sous pli cacheté, un dispositif permettant d'enregistrer et de restituer les sons, fortement inspiré du principe du *Phonautographe*. Les vibrations du diaphragme ne sont plus inscrites sur un papier mais gravées par une aiguille sur un disque tournant. Pour la restitution, les gravures font vibrer le diaphragme dont le son est amplifié par le cornet. Ce dispositif n'existe que dans l'esprit de Cros et n'a pas fait l'objet d'un début de réalisation concrète. Quatre mois plus tard, l'américain Thomas Edison présente son *Phonographe* qui, lui, est fonctionnel et stupéfie le monde entier. Le message enregistré est très distordu, nasillard, bruyant et de faible niveau mais intelligible. Ce dispositif est très proche de celui décrit par Cros. Il n'en diffère que par la gravure en profondeur selon un sillon hélicoïdal, sur une plaque d'étain disposée sur un cylindre de 10 cm de diamètre tournant à 1 tour par seconde. Bien que le *Phonographe* suscite alors un vif intérêt, il présente trois désavantages importants : 1/ le son enregistré sur le cylindre ne peut être reproduit convenablement qu'une seule fois, car la reproduction en détériore la surface. 2/ il faut tourner régulièrement la manivelle pendant tout l'enregistrement et la reproduction pour assurer la rotation continue du cylindre, ce qui représente une tâche fatigante. 3/ enfin le son enregistré sur le cylindre est un original qui ne peut pas être copié. Le *Phonographe* ne fut pour cela considéré pendant longtemps que comme une curiosité, et il fallut une dizaine d'années pour que des perfectionnements importants lui soient apportés par Edison et bien d'autres inventeurs pour en permettre une utilisation institutionnelle. L'utilisation de cire puis de gomme-laque comme surface de gravure permit une amélioration très importante de la qualité sonore de la restitution qui de surcroît, grâce à l'amélioration des aiguilles de lecture, put être reproduite sans perte de qualité. L'entraînement du cylindre fut assuré par un moteur mécanique, en revanche, il ne put jamais être dupliqué malgré tous les efforts d'Edison. Ainsi débarrassé de ses défauts de jeunesse, le *Phonographe* se présente dès 1886 comme un instrument incontournable pour les chercheurs travaillant sur la parole. Mais si celui-ci enregistre et restitue les sons, leurs tracés sont tout à fait différents que ceux donnés par les divers *Phonautographes*. En effet, ces derniers donnent une représentation des variations d'amplitude de la pression acoustique en fonction du temps comme nous la connaissons de nos jours alors que les sillons du *Phonographe* gravés en profondeur reproduisent bien le son mais en donnent une image totalement différente. Cependant, ces images des sillons ont un grand avantage car leur fidélité est attestée par l'écoute, qui donne la preuve de l'objectivité du signal acoustique qu'elles représentent.

Le *Phonographe* va très vite trouver un concurrent redoutable dans le *Gramophone* proposé par l'Allemand Emil Berliner en 1888 qui ne cache pas avoir été très inspiré par les idées de Cros. Le cylindre est remplacé par un disque dont les gravures des sillons sont latérales et non en profondeur. Grâce à ce support, le *Gramophone* est mieux adapté à la duplication par pressage et à la diffusion de programmes musicaux et il va bientôt dominer le marché. Cependant, sa fonction d'enregistrement est moins bonne que celle du *Phonographe* et ce dernier va rester l'outil préféré des phonéticiens et acousticiens jusqu'à l'apparition des enregistreurs-lecteurs sur disques souples vers 1930.

4. LES IMPASSES

Parallèlement au développement des divers *Phonotographes*, d'autres techniques ont été imaginées pour rendre le signal de parole visible avec plus ou moins de bonheur.

4.1. Les flammes manométriques

La méthode des *flammes manométriques* a été découverte un peu par hasard par Rudolph Koenig en 1882. Elle est basée sur la constatation qu'un son peut moduler l'amplitude et la forme de la flamme d'un bec alimenté en gaz de ville par l'intermédiaire d'une capsule manométrique. On observe les flammes grâce à un miroir tournant mais on ne peut pas en fixer les images qui sont très fugaces. La forme des flammes donne une représentation de la structure acoustique des voyelles qui permet de les différencier. Ainsi, dans la description de chaque période on peut distinguer de une à quatre flammes de différentes largeurs et amplitudes, dont le positionnement dans l'image du cycle périodique varie en fonction du timbre du son (figure 4). Ces flammes correspondent à la structure harmonique du spectre. Elles ont été un peu utilisées pour l'étude des voyelles mais surtout pour leur qualité pédagogique. René Marage, un élève d'Etienne-Jules Marey, réussit en 1895 à fixer l'image des flammes sur un film et étudia ainsi des phénomènes de filtrage par des tubes et des cornets qui sont cohérents avec les connaissances actuelles. Ce furent leur dernière application.

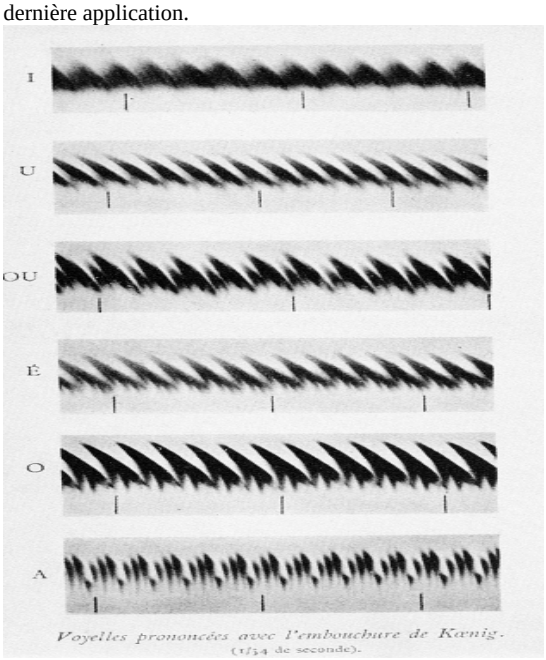


Figure 4 : Représentation du signal acoustique des voyelles au moyen de chronophotographies des flammes de Koenig par Marage en 1896 (d'après Marey, 1898)

4.2. Les tracés directs du phonographe

La gravure en profondeur sur les cylindres de cire du *Phonographe* donne des images du signal totalement différentes de celles des *Phonautographes*. Malgré ce désavantage, certains chercheurs s'adaptèrent à cette nouvelle représentation du signal de parole et tentèrent de la codifier. Les gravures étant très petites, leur étude nécessita l'utilisation d'un microscope ou d'agrandissements photographiques et s'avèra fastidieuse. L'attention de l'observateur étudiant un sillon de *Phonographe* se porte sur la forme de la gravure qui définit le timbre du son, sur la profondeur qui définit son intensité et sur la périodicité des segments vocaliques qui définit sa hauteur (figure 5). Si ce type de description est peu utile pour comparer les timbres vocaliques en raison de la trop grande complexité morphologique des sillons, il permet en revanche des mesures précises de durées. Hector Marichelle, un autre élève de Marey, semble avoir poussé le plus loin la lecture directe des sillons du *Phonographe* grâce à laquelle il a mené les premières études sur la prosodie de la parole en 1896. Après lui, la lecture directe des sillons sera remplacée par la transcription gravure-écriture.

4.3. Le téléphone écrivant

En 1876, Bell invente le microphone électromagnétique. Avant même qu'il ne propose le *Téléphone* deux ans plus tard, de nombreux chercheurs tentèrent de l'adapter à la méthode graphique pour tracer le signal de parole sur les cylindres enduits de noir de fumée. En 1882, le Français Boudet de Parîs proposa le *Téléphone écrivant* suivi de près par Pierre-Jean Rousselot (l'Abbé). Comme pour les autres tentatives, ces instruments furent décevants face aux dernières versions des *Phonautographes*, essentiellement à cause de leur système d'inscription. L'*inscripteur électrique* ne se développa que bien plus tard après l'apparition de l'électronique avec l'invention de la triode.

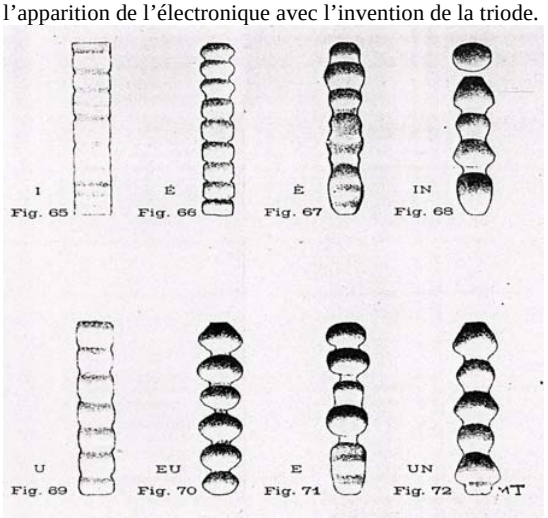


Figure 5 : Empreintes de sillons gravés dans un cylindre de cire de phonographe représentant une période de plusieurs voyelles par Marichelle en 1897 (d'après Marey, 1898)

5. LES TAMBOURS DE MAREY

Etienne-Jules Marey est un physiologiste auquel la communauté scientifique doit l'essentiel des méthodes d'exploration du mouvement dont la *méthode graphique*, pour laquelle il a développé un système pneumatique de transcription d'un grand nombre de phénomènes : les fameux tambours à leviers qui portent son nom. A la suite de ses désillusions avec l'inscripteur électrique, Rousselot, qui va devenir la référence quasi universelle de la phonétique expérimentale, découvre en 1888 que les tambours de Marey sont capables, dans certaines conditions, de tracer les vibrations de la voix sur les cylindres enduits de noir de fumée. Il ne cessera ensuite d'améliorer l'inscription du signal vocal au moyen des tambours par des travaux empiriques sur les membranes pour en augmenter les performances en terme de bande passante en fréquence et de sensibilité. Ainsi ses « petits tambours », associés à une bonne loupe et à la

photographie pour en agrandir les tracés, vont lui permettre de distinguer les différentes voyelles (figure 6) et il en préconisera toujours l'utilisation en association avec des systèmes d'inscription plus modernes et de fait ils seront l'outil privilégié des phonéticiens pendant plus d'un demi-siècle. Car les tambours de la *méthode graphique* ont un double intérêt, d'une part, il est possible d'enregistrer simultanément et en synchronie d'autres paramètres, tels que physiologiques, sur le même support. D'autre part, la trace du signal peut être analysée mathématiquement après agrandissement au moyen des séries de Fourier pour en déduire sa structure harmonique. Par contre, la fidélité de la trace du signal est difficile à obtenir et surtout à prouver.

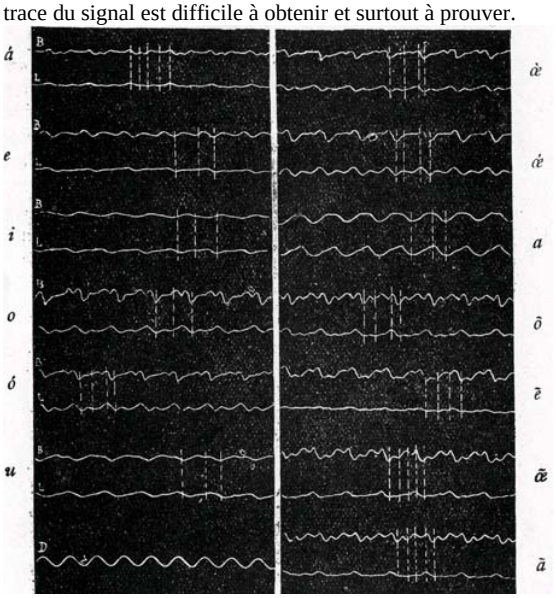


Figure 6: Traces du signal du phonogramme buccal (B) et des vibrations du larynx (L) enregistrées avec des tambours à levier pour différentes voyelles. Base de temps (D) à 200 Hertz (d'après Rousselot, 1897).

6. LA TRANSCRIPTION GRAVURE-ECRITURE

A partir de la fin des années 1880, le *Phonographe* est devenu un instrument indispensable aux chercheurs du domaine de la parole. Pour pallier l'incompatibilité de sa trace avec celle de la *méthode graphique*, de multiples tentatives de transcription gravure-écriture vont donc être menées dans le but d'unifier les avantages des deux méthodes et de permettre de *« voir et entendre simultanément »* le signal de parole. Elles sont toutes basées sur des méthodes mécano-optiques et utilisent la possibilité de ralentir le signal en diminuant la vitesse de rotation des cylindres en lecture. La conversion des gravures en profondeur des sillons du *Phonographe* en représentation temps-amplitude des tracés de la *méthode graphique* auxquels nous sommes tous habitués a toujours été problématique et n'a jamais été techniquement résolue. C'est l'Allemand Ludimar Hermann qui, à partir de 1893, en a été le premier animateur. Son système était inspiré du principe du *Phonautographe* à transcription optique de Blake dont le miroir était directement mu par l'aiguille lectrice du *Phonographe*. Bien d'autres dispositifs dont l'efficacité et la pratique n'étaient pas très bien assurées ont été proposés, dont des solutions purement mécaniques telles que le transcripteur de l'Américain Edward Scripture qui convertissait en 1906 les gravures en profondeur des cylindres du *Phonographe* en gravures latérales sur des disques de *Gramophone* et surtout le Français Théodore Rosset, le fondateur, voici un siècle, de l'Institut de Phonétique de Grenoble. Ce dernier réalisa en 1911 un transcripteur très astucieux entre deux cylindres, l'un émetteur à gravure en profondeur et l'autre récepteur à gravures latérales. Un enregistreur optique sur papier photographique permettait de contrôler la qualité de la transcription. Mais ces convertisseurs furent toujours l'objet de polémiques entre phonéticiens. La plus fameuse, entre Rousselot et Rosset ne cessera qu'à la disparition des protagonistes.

Mais à la même époque, Valdemar Poulsen avait déjà inventé le principe du magnétophone et Lee de Forest la triode.

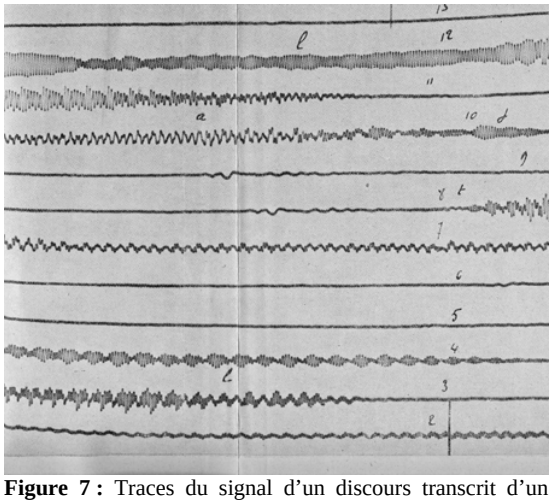


Figure 7 : Traces du signal d'un discours transcrit d'un cylindre de *Phonographe* avec le transcripteur de Rosset (d'après Rosset 1911).

L'électronique allait gagner tous les champs de l'instrumentation scientifique, et ce sont ses propres progrès qui vont rythmer la course à la trace du signal de parole, qui va continuer à travers l'évolution des microphones, oscilloscopes et enregistreurs graphiques jusqu'aux éditeurs de signaux. Mais ceci est une autre histoire.

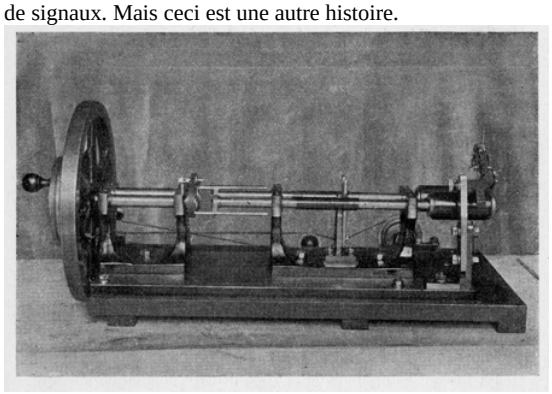


Figure **8** : Transcripteur mécanique de Rosset (d'après Rosset, 1911).

BIBLIOGRAPHIE

Marey, E. J., 1898 « L'inscription des phénomènes phonétiques. Première partie : Méthodes directes », *Revue générale des sciences pures et appliquées*, n°11, p. 445-456.

Rosset, T., 1911, *L'inscription de la voix parlée*, Armand Colin, Paris, 102 p.

Rousselot, P. J., 1897, *Principes de Phonétique expérimentale*, t. 1, Welter, Paris, 638 p.

AUGMENTATION DU TAUX DE FAUSSE ACCEPTATION PAR TRANSFORMATION INAUDIBLE DE LA VOIX DES IMPOSTEURS

Jean-François Bonastre, Driss Matrouf, Corinne Fredouille

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{jean-francois.bonastre,driss.matrouf,corinne.fredouille}@univ-avignon.fr

ABSTRACT

This paper investigates the effect of a transfer function-based voice transformation on automatic speaker recognition system performance. We focus on increasing the impostor acceptance rate, by modifying the voice of an impostor in order to target a specific speaker. This paper is based on the following idea : in several applications and particularly in forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used and could impersonate a given, well known speaker. We also evaluate the effect of the voice transformation when the transformation is applied both on client and impostor trials.

1. INTRODUCTION

La voix est une modalité biométrique compétitive pour plusieurs raisons bien connues. En particulier, cette modalité est souvent la seule disponible pour de nombreux types d'applications. Malgré une fiabilité largement inférieure comparée - par exemple - à l'iris, les progrès enregistrés dans les dernières décennies ont mené à l'émergence de systèmes automatiques utilisables dans des applications commerciales. Cependant, les nombreux facteurs de variabilité intervenant dans cette modalité, difficiles à contrôler et à prévoir, forment une limite aux performances des systèmes. Les différences dues au microphone, à l'environnement et aux différences inter-sessions sont les facteurs les plus présents dans la littérature, pour leur importance d'une part, mais également pour leur plus grande fréquence.
Durant la même période, dans le champ de la criminalistique, les juges, les avocats, les enquêteurs et les agences de sécurité nationale se sont montrés très demandeurs de techniques permettant d'identifier un individu par sa voix, pour confondre un suspect ou pour servir d'élément de preuve dans un tribunal [1]. En dépit du fait que la communauté scientifique, dans une large majorité, ait remis en cause les bases scientifiques d'une telle identification vocale [2][3][4] et du message fort de "besoin de précaution" émis par [5], les techniques d'identification vocale sont couramment utilisées dans la pratique criminalistique, particulièrement dans le contexte d'événements terroristes à l'échelle mondiale, avec l'utilisation de plus en plus fréquente de systèmes automatiques.
Ce papier vise à identifier une des limites des systèmes automatiques, ou plutôt de leur utilisation dans ce contexte spécifique : si vous connaissez la technique de reconnaissance utilisée pour identifier une voix, si vous avez également un exemple de la voix d'une personne X, est-il possible de transformer la voix d'une personne Y de telle façon que le système conclut à une identité entre X et Y, sans que cette transformation ne soit audible ?
Les objectifs visés dans ce papier sont proches de l'approche "voice-forgery" proposée dans [16], la différence principale réside dans l'interprétation de la transformation : dans notre cas, le système de reconnaissance doit être trompé quand le but est de synthétiser une voix proche de X au sens perceptuel pour la "voice-forgery". Ce papier étend les travaux préliminaires présentés dans [10][11].

La section 2 présente l'approche statistique couramment utilisée en reconnaissance du locuteur et dans laquelle s'inscrit ce travail. La section 3 décrit la technique de transformation de voix que nous utilisons. La section 4 décrit le protocole expérimental utilisé et les résultats sont présentés dans la section 5. Enfin, la section 6 conclut ce travail et ouvre quelques perspectives.

2. L'APPROCHE GMM-UBM EN RECONNAISSANCE DU LOCUTEUR

L'approche GMM-UBM (Gaussian Mixture Model - Universal Background Model) est la technique prédominante en reconnaissance du locuteur, en mode indépendant du texte [9]. Etant donné un segment de parole Y et un locuteur S, la tâche de vérification du locuteur consiste à déterminer si Y a été prononcé par S. Cette prise de décision est modélisée par l'estimation d'un rapport de deux probabilités : Y provient de S ($H0$) et Y a été prononcé par un inconnu ($H1$). Ce rapport (LR, pour Likelihood Ratio) est comparé à un seuil de décision θ ; $H0$ est désignée si le ratio est supérieur au seuil, $H1$ dans le cas contraire. De manière plus formelle, le ratio LR est donné par :

$$LR(Y, H0, H1) = \frac{p(Y|H0)}{p(Y|H1)} \quad (1)$$

avec Y, le segment de parole à tester, $p(Y|H0)$, la vraisemblance de $H0$, $p(Y|H1)$ la vraisemblance de $H1$ et θ, le seuil de décision.
Un modèle, λ_{hyp} représente l'hypothèse $H0$; ce modèle est appris à partir d'un exemple de la voix du locuteur concerné. Le modèle $\lambda_{\overline{hyp}}$ représente la seconde hypothèse, $H1$, et est généralement appris à partir d'une collection d'extraits vocaux provenant d'un grand ensemble de locuteurs.
Le ratio LR devient $\frac{p(Y|\lambda_{hyp})}{p(Y|\lambda_{\overline{hyp}})}$. Les deux modèles sont des modèles à mélange de lois gaussiennes (GMM) :

$$p(x|\lambda) = \Sigma_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) \quad (2)$$

avec w_i, μ_i et Σ_i, les poids, les vecteurs moyennes et les matrices de covariance (en général diagonales) des différentes composantes

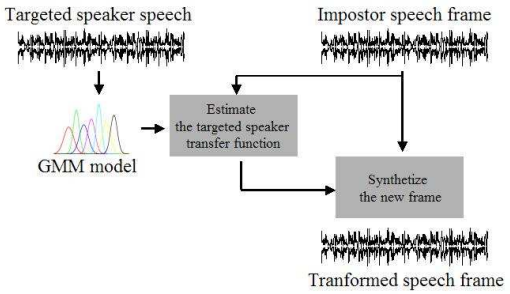


Fig. 1. Principe de la transformation, pour une trame de signal.

du mélange. Le modèle $\lambda_{\overline{hyp}}$ est dénoté modèle du monde, ou UBM quand il est indépendant de l'environnement et est estimé en maximisant la vraisemblance des données d'apprentissage correspondante. Le modèle λ_{hyp} est dérivé de l'UBM, par une technique de maximisation de la probabilité $aposteriori$ (MAP). De manière courante, seules les moyennes des composantes sont adaptées et les autres paramètres sont directement issus de l'UBM [12].

3. TRANSFORMATION DES VOIX

L'objectif de la transformation de voix présentée dans cette section est d'augmenter la vraisemblance d'un signal Y étant donné un locuteur X, ou plus exactement étant donné le GMM représentant ce locuteur. La voix transformée doit conserver un aspect naturel pour un auditeur.
Le principe retenu pour cette transformation consiste à analyser le signal Y trame à trame en utilisant un modèle source-filtre et à modifier la fonction de transfert pour la rapprocher du locuteur cible. Celui-ci est modélisé par un modèle GMM, appris sur un extrait de parole lui appartenant ; ce modèle servant à la transformation du modèle du locuteur cible. En fait, la fonction de transfert initiale est remplacée par une fonction de transfert estimée par une moyenne arithmétique des moyennes des différentes composantes du modèle cible, pondérées par leur probabilité $aposteriori$ pour la trame considérée. Le signal est enfin reconstitué par une simple approche "overlap-add". La figure 1 présente une vue générale du procédé de transformation pour une trame du signal Y.
En réalité, le procédé est un peu plus complexe, par le fait que chaque système de reconnaissance du locuteur utilise une représentation spécifique des données acoustiques, issue d'une étape de calcul de coefficients acoustiques suivie de différents étages de normalisation (de manière classique : sélection des trames de parole puis centrage et réduction des données sélectionnées) rendant parfois impossible l'estimation de la fonction de transfert cible. Pour contrer ce problème, nous avons appris deux modèles pour le locuteur ciblé : un modèle en respectant la représentation des données du système de reconnaissance du locuteur et un modèle utilisant une représentation des données plus simple, propre au système de transformation de voix. Ces deux modèles sont appris en parallèle, le premier guidant le second (i.e. les probabilités $aposteriori$ pour qu'une donnée soit liée à une composante du modèle - utilisées au sein de l'algorithme d'estimation EM/MAP - sont issues du premier modèle).

Soit y, une trame du signal à transformer, provenant d'un imposteur S' et x, la trame correspondante, appartenant au locuteur cible, S . Le modèle source-filtre amène :

$$Y(f) = H_y(f)S_y(f) \quad (3)$$
$$X(f) = H_x(f)S_x(f) \quad (4)$$

avec Y et X, les représentations spectrales de y and x, H_y et H_x, les fonctions de transfert correspondant respectivement à y et x, S_x et S_y, les transformées de Fourier du signal de la source pour x et y (il faut noter que le spectre n'est rien d'autre qu'une représentation compacte de la fonction de transfert). Pour rapprocher y de x - en termes de forme spectrale - il est suffisant de remplacer H_y par H_x dans l'équation 3 :

$$Y'(f) = H_x(f)S_y(f) = \frac{H_x(f)}{H_y(f)}Y(f) \quad (5)$$

Pour réaliser cela, si nous décidons de ne pas modifier la phase du signal original, nous appliquons le filtre suivant au signal y :

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \quad (6)$$

Dans ce processus de filtrage, pour chaque trame y (imposteur) nous avons besoin de connaitre H_x (et non x voir equation 6). Pour des commodités statistiques, nous préférons estimer une version LPCC de H_x que nous dénommons $Hlpcc_x$ (le passage de l'une à l'autre est trivial [11]) : $Hlpcc_x = Sigma_g[P(g|y) * m_{g,lpcc}]$. Les probabilités $P(g|y)$ sont calculées en utilisant le gmm cepstral (dans le domaine de représentation des données du système de verification du locuteur) du client. Les $m_{g,lpcc}$ sont les moyennes des composantes du modèle de ce client dans le domaine LPCC (la correspondance entre les composantes des deux modèles est maintenue lors de l'apprentissage de ceux-ci).

4. PROTOCOLE EXPÉRIMENTAL

Les expériences permettant de valider l'approche choisie dans ce papier ont été réalisées sur la base du protocole de la campagne d'évaluation NIST-SRE 2005 [13].
Le corpus utilisé correspond au corpus de NIST-SRE 2005, réduit à la partie "homme" de la tâche "one-conv/one-conv". Les messages vocaux utilisés pour l'apprentissage des modèles des clients et pour les tests sont d'une durée moyenne de 2mn30 de parole téléphonique et conversationnelle. Le protocole comporte 1231 tests "clients" et 12317 tests "imposteurs". Le modèle UBM ainsi que les données nécessaires à la normalisation des scores sont issus des corpus des campagnes SRE des années 2002 à 2004.
Trois expériences ont été réalisées :

- Une expérience de calibration, réalisée sans utiliser la transformation de voix (baseline).
- Une expérience où la transformation de voix est appliquée pour chaque test "imposteur", avec la connaissance du message vocal utilisé pour entraîner le modèle du locuteur ciblé (expérience 1).
- Une expérience où la transformation de voix est appliquée pour chaque test "imposteur", en utilisant un message vocal venant du locuteur ciblé, mais différent du message utilisé

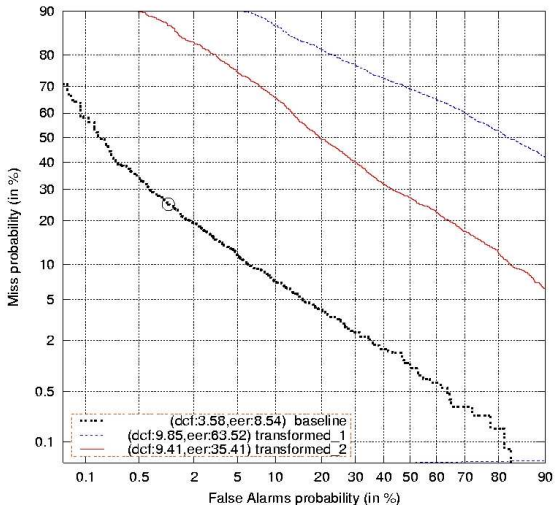


Fig. 2. Courbes DET pour le baseline, l'expérience 1 (utilisant les données d'entraînement pour la transformation des voix) et l'expérience 2 (utilisant un enregistrement différent pour la transformation de voix).

pour l'apprentissage, excepté pour un faible nombre de locuteurs, pour lesquels un seul enregistrement est disponible (expérience 2).

Pour toutes les expériences, le même modèle du monde a été utilisé, pour le procédé de transformation de voix comme pour les tests de reconnaissance.

Nous avons utilisé le système LIA_SpkDet [14], développé au LIA sur la base du toolkit ALIZE [17][15]. L'ensemble des logiciels utilisés est disponible sous forme de logiciels libres.
Le système LIA_SpkDet est basé sur l'approche UBM-GMM et implémente une normalisation des scores de type TNORM. La paramétrisation acoustique correspond à 16 coefficients cepstraux LFCC augmentés des 16 dérivées premières. Une sélection des vecteurs acoustiques, basée sur une modélisation multi-gaussienne de l'énergie est réalisée, avant de centrer et de réduire les coefficients. Le modèle UBM ainsi que les modèles de locuteurs comportent 2048 composantes. Durant les tests, les 10 meilleures composantes sont sélectionnées et utilisées.

5. RÉSULTATS

L'influence du procédé de transformation de voix est mesurée en termes de courbes DET, présentant les taux d'erreur de type I en fonction du taux d'erreur de type II, et à travers les distributions des scores imposteurs. La figure 2 présente les courbes DET pour le baseline, l'expérience 1, où les voix des imposteurs ont été transformées en utilisant les données d'apprentissage des locuteurs ciblés et l'expérience 2, pour laquelle des données différentes ont été employées pendant la transformation. La normalisation TNORM a été employée dans tous les cas. La figure 3 montre les distributions des scores imposteurs pour les trois expériences. Les figures 4 et 5 montrent les distributions des scores clients et imposteurs du

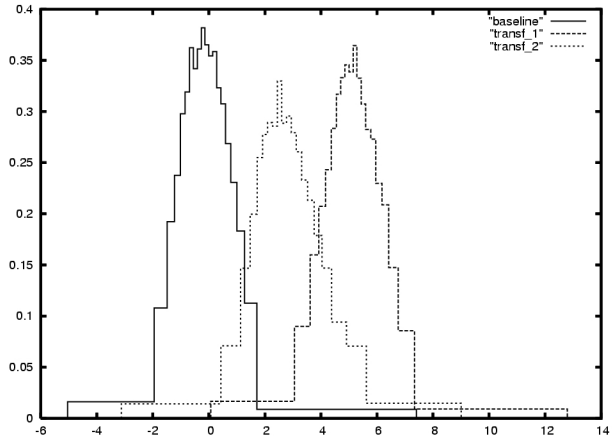


Fig. 3. Distribution des scores imposteurs pour le baseline (baseline), l'expérience 1 (transf_1) et l'expérience 2 (transf_2)

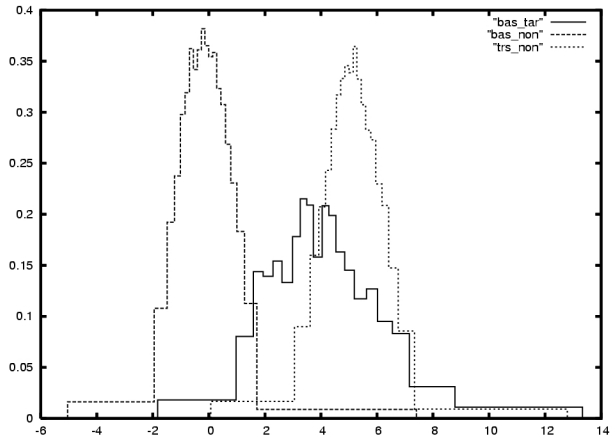


Fig. 4. Distributions des scores clients pour le baseline (bas_tar) et des scores imposteurs pour le baseline (bas_non) et pour l'expérience 1 (trs_non).

baseline, comparées respectivement à la distribution imposteur de l'expérience 1 et de l'expérience 2.
Un important déplacement de la distribution des scores imposteurs est clairement mis en évidence pour les deux expériences où la transformation de voix est appliquée, comparé au baseline. Le même effet est observé sur les courbes DET, qui montrent une très forte dégradation des performances. Bien entendu, l'expérience 1, qui utilise directement les données d'apprentissage pour la transformation, présente une dégradation plus importante que l'expérience 2, plus réaliste, dans laquelle un enregistrement différent de celui de l'apprentissage est utilisé.

Pour mettre en évidence l'évolution du taux de fausses alarmes, nous proposons dans la table 1 les taux de faux rejet et de fausse alarme pour un seuil $apriori$, fixé empiriquement sur un ensemble de tests différent. L'influence de la transformation de voix sur le taux de fausse alarme est clairement mise en évidence.

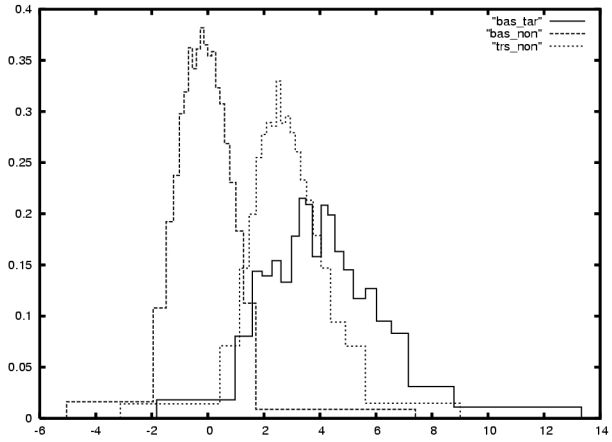


Fig. 5. Distributions des scores clients pour le baseline (bas_tar) et des scores imposteurs pour le baseline (bas_non) et pour l'expérience 2 (trs_non).

	False Alarm (%)	Miss Probability (%)
Baseline	0.88	27.45
Exp 1	96.55	27.45
Exp 2	49.72	27.45

Table 1. Taux de "False Alarm" et de "Miss Probability" en utilisant un seuil défini $apriori$, pour les 3 expériences.

6. CONCLUSION ET PERSPECTIVES

Dans ce papier, nous avons évalué les effets d'une transformation artificielle de la voix sur les taux de fausse alarme d'un système de reconnaissance du locuteur. L'objectif principal était de vérifier si il était possible de tromper un tel système lorsqu'on possède un exemple de la voix d'un locuteur cible et des connaissances sur le système utilisé. Sous ces hypothèses, qui semblent tout à fait réalistes par exemple dans le contexte d'événements terroristes de grande ampleur, un procédé très simple de transformation de la voix a permis, sans dénaturer la voix, de perturber très sensiblement le système de reconnaissance utilisé. L'approche et les résultats proposés mettent en lumière une des limites des approches utilisées en reconnaissance automatique du locuteur, dans le cadre criminalistique.
Dans cette étude, une large connaissance du système de reconnaissance du locuteur employé était utilisée pour la transformation de voix : technique de base, paramétrisation et modèle du monde notamment. Une première continuation de ce travail consistera donc à mesurer l'influence des différents éléments cités précédemment. Des améliorations du procédé de transformation sont également envisagées, en lissant les transformations par exemple.

7. REFERENCES

[1] R.H. Bolt, F.S. Cooper, D.M. Green, S.L. Hamlet, J.G. McKnight, J.M. Pickett, O. Tosi, B.D. Underwood, D.L. Hogan, "On the Theory and Practice of Voice Identification", *National Research Council, National Academy of Sciences, Washington, D.C.*, 1979.

[2] R.H. Bolt, F.S. Cooper, E.E.Jr. David, P.B. Denes, J.M. Pickett, K.N. Stevens, "Speaker Identification by Speech Spectrograms : A Scientists' View of its Reliability for Legal Purposes",*Journal of the Acoustical Society of America*, 47, 2 (2), 597-612, 1970.

[3] J.F. Nolan, "The Phonetic Bases of Speaker Recognition",*Cambridge University Press : Cambridge*, 1983.

[4] L.J. Boë, "Forensic voice identification in France",*Speech Communication, Elsevier*, Volume 31, Issues 2-3, June 2000, pp. 205-224 (http ://dx.doi.org/10.1016/S0167-6393(99)00079-5).

[5] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.P. Campbell, D.A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice : A Need for Caution", *Proceeding of Eurospeech 2003*, 2003

[6] C. Champod, D. Meuwly, "The inference of identity in forensic speaker recognition",*Speech Communication*, Vol. 31, 2-3, pp 193-203, 2000

[7] J. González-Rodríguez, J. Ortega, and J.J. Lucena, "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", *Proc. Odyssey 2001 Speaker Recognition Workshop*, pp. 135-138, Crete (Greece), 2001

[8] P. Rose, T. Osanai, Y. Kinoshita, "Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Speech Language and the Law*, 2003 ; 10/2 : 179-202.

[9] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 2004, Vol.4, pp.430-451

[10] D. Matrouf, J.-F. Bonastre and J.P. Costa, "Effect of impostor speech transformation on automatic speaker recognition", *proc. of COST 275 Workshop "Biometric on the internet"*, Hatfied, UK, 2005

[11] D. Matrouf, J.-F. Bonastre and C. Fredouille, "Effect of voice transformation on impostor acceptance", *Proc. of ICASSP 2006*, Toulouse, France, 2006

[12] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing (DSP), a review journal Special issue on NIST 1999 speaker recognition workshop*, vol. 10(1-3), pp 19-41, 2000.

[13] NIST Speaker Recognition Evaluation campaigns web site, http ://www.nist.gov/speech/tests/spk/index.htm

[14] LIA_SpkDet system web site, http ://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[15] J.-F. Bonastre, F. Wils, S. Meignier, "ALIZE, a free toolkit for speaker recognition", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[16] P. Perrot, G. Aversano, R. Blouet, M. Charbit, G. Chollet, "Voice Forgery Using ALISP : Indexation in a Client Memory", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[17] ALIZE project web site, http ://www.lia.univ-avignon.fr/heberges/ALIZE/

Session II

Conférence Invitée

Lundi 12 juin 2006 - 14h30 15h30

Le langage humain à la lumière de l'évolution

Jean-Louis Dessalles

ParisTech
École Nationale Supérieure des Télécommunications
46 rue Barrault – F-75013 Paris, France
dessalles@enst.fr
www.enst.fr/~jld

ABSTRACT

This paper explores a few consequences of the hypothesis that language evolved for the benefit of speakers. The hypothesis, supported by recent Darwinian scenarios of language emergence, explains why speech production organs were dramatically transformed through evolution, while auditory systems remained practically unchanged. It also explains the need for huge vocabularies and for large episodic memory, and it dismisses the possibility of gesture-first scenarios of language origins.

1. INTRODUCTION

L'étude scientifique des propriétés caractéristiques des espèces animales ne se conçoit pas en dehors de la recherche de leur fonction biologique. Les chercheurs qui étudient la coordination des milliers de muscles de la trompe des proboscidiens ou le comportement bâtisseur de *Castor canadensis* ont une idée de l'avantage que ces animaux tirent de leur particularité. De manière étonnante, l'étude scientifique du langage a longtemps été menée en dehors de tout questionnement sur la fonction biologique de cette particularité humaine. Se poser la question : *À quoi sert le langage ?* permet pourtant de répondre à certaines interrogations et de soulever de nouveaux problèmes, autrement dit de faire progresser notre connaissance de cette faculté unique. Ce papier vise à le démontrer à l'aide de quelques exemples concrets.

Pour des raisons historiques liées à la confusion avec l'origine des langues [29], puis parce que l'on a cru que la faculté de langage était apparue indépendamment de toute pression de sélection ([9], p. 75), la question de la raison d'être du comportement langagier a longtemps été évacuée de l'investigation scientifique. On peut dater le renouveau des recherches sur l'origine du langage de la publication par Steven Pinker et Paul Bloom d'un article [33] qui est parvenu à replacer la question des origines du langage dans un cadre darwinien. Depuis, un nombre important d'ouvrages et d'articles ont abordé cette question, qui s'est montrée d'une richesse insoupçonnée.

Dans ce qui suit, nous exposerons tout d'abord le cadre théorique dans lequel nous nous plaçons, qui est un cadre strictement darwinien. Il nous permettra de conclure que, de manière paradoxale, le langage bénéficie surtout à celui qui parle. Nous en tirerons ensuite un certain nombre de conséquences concernant le caractère référentiel du langage, la taille disproportionnée du lexique qui justifie une phonologie digitale, la raison d'être de la mémoire épisodique et le caractère universellement oral des langues, malgré la supériorité potentielle du geste comme vecteur d'information.

L'INTÉRÊT BIOLOGIQUE DE LA PAROLE

Une attitude préscientifique répandue consiste à considérer que le langage existe parce qu'il permet à l'espèce qui en est dotée d'échanger des informations utiles sur son environnement. Une telle explication n'est pas recevable dans un cadre darwinien. Les être vivants actuels descendent d'individus qui se sont reproduits davantage, non pas que les individus d'autres espèces, mais que leurs propres congénères. La sélection naturelle résulte d'une compétition intra- et non inter-espèces. Autrement dit, le succès écologique d'une espèce n'a aucune valeur de prédiction quant à son évolution [43].

Les calculs de théorie des jeux appliqués à la théorie de la sélection naturelle [27] ont montré qu'un comportement qui n'est pas dirigé vers des individus apparentés se doit de bénéficier à son auteur pour s'imposer. Dans le cas du langage, la contrainte ne va pas soi. Si le langage véhicule des informations potentiellement utiles, le bénéfice de la prise de parole semble aller intégralement aux auditeurs et non au locuteur, ce qui pose un problème considérable dans un contexte darwinien.

Plusieurs tentatives théoriques ont été faites pour sortir de ce paradoxe. Certaines visent à soustraire le langage aux lois de la nature en en faisant un pur produit des habitudes culturelles ([30], p. 214 ; [39], pp. 94 & 208). Une telle position est difficilement tenable au regard de données comme la position 'anormalement' basse de notre pharynx, qui ne se justifie que s'il s'agit d'une adaptation biologique au langage. Un autre phénomène incompatible avec la théorie purement culturelle nous est donné par l'existence d'universaux langagiers. Citons notamment (voir [16]) :

- la phonologie digitale
- la récursivité centrale dans la syntaxe
- le blocage de la coréférence quant un pronom lie un nom (ex. *elle dit que la sœur de Leïla est malade*)
- les lois de la narration (sensibilité à la proximité, à la récence, à l'improbabilité…)

- les lois de l'argumentation (le mécanisme contradiction–abduction–négation)

Ces phénomènes et d'autres, tous spécifique au langage, sont trop profondément ancrés dans la cognition pour qu'on puisse facilement attribuer leur universalité à un hypothétique héritage culturel. Si l'on reconnaît donc un ancrage biologique au langage, le problème revient dans toute son acuité : comment la prise de parole peut-elle profiter au locuteur ?

Les hypothèses avancées pour tenter de résoudre ce problème ne sont pas nombreuses. L'une consiste à voir dans le langage un échange coopératif d'informations utiles, une sorte de troc ([35], p. 28 ; [42] ; [5]). Ainsi, la prise de parole bénéficierait à l'auditeur dans un premier temps, mais également au locuteur dans un second temps lorsque les rôles s'inversent. Malheureusement, cette manière de voir pose de nombreux problèmes [15]. La prise de parole, dans un scénario coopératif, devrait être toujours utilitaire, parcimonieuse et dirigée vers un seul interlocuteur à la fois, de manière à éviter le vol d'information qui ruinerait l'avantage du locuteur. Or, la réalité du langage est bien différente : les prises de paroles de la vie quotidienne ne portent pas systématiquement sur des informations réellement utiles ; il existe plus d'individus bavards en quête de public que de détenteurs d'informations attendant d'être sollicités ; et les individus s'adressent souvent à plusieurs personnes, près de deux en moyenne [20].

La seule hypothèse alternative actuellement disponible est que le langage constitue un moyen d'affichage, ce qui permet de le faire entrer dans le cadre théorique du signal honnête (*honest* ou *costly signalling*) [23], qui est un cas particulier de la théorie du handicap [44]. Cette perspective confère une fonction fondamentale au langage, celle de participer à l'établissement des liens sociaux [21]. Les individus parleraient pour afficher certaines qualités, notamment *la capacité à savoir avant les autres*, qui sont recherchées par les partenaires de coalitions [16]. L'avantage du parleur, dans ce cas, est immédiat : il est de s'attirer ou de conserver les meilleurs alliés auxquels il peut prétendre [19].

Les conséquences d'un tel scénario sont multiples et importantes, comme nous le verrons dans le reste de cet article. La toute première conséquence est que *les pressions sélectives se sont exercées essentiellement sur la production de la parole*, puisque ce sont les locuteurs qui cherchent à afficher certaines qualités. Pour prendre une métaphore, ce sont les marchands font l'effort de dresser un étale et de vanter leur marchandise, tandis que les clients continuent à avoir un comportement normal, se contentant tout au plus de marchander. On s'attend donc à ce que l'essentiel des transformations physiologiques et comportementales liées au langage se situent du côté de la production des sons et des énoncés.

LA PRODUCTION DES SONS

La communication langagière humaine repose sur un mécanisme vocal de production et un mécanisme acoustique d'analyse. D'un point de vue évolutif, la dissymétrie entre les deux est patente. Le système de production des sons s'est considérablement transformé par rapport à ce qu'il est chez les autres primates, alors que les capacités d'audition sont pratiquement restées inchangées, en parfaite conformité avec l'idée que le langage bénéficie essentiellement au locuteur.

L'appareil vocal humain a évolué pour augmenter significativement la cavité résonante constituée par le pharynx, qui est descendu au niveau de la sixième vertèbre, alors qu'il est positionné au niveau de la deuxième vertèbre chez d'autres mammifères comme le chien [4]. Cette position basse pose des problèmes importants, notamment les nombreux accidents liés à l'aspiration de nourriture.[1] Une autre spécificité humaine concerne le contrôle cortical direct de l'appareil vocal, qui offre une commande volontaire précise des muscles laryngés, absente chez les autres primates ([14], p. 250).

Le contraste est saisissant avec les capacités acoustiques, qui n'ont pas connu de changement spectaculaire si on les compare avec celles des primates ou même avec celles des mammifères et des oiseaux. La capacité à discriminer les phonèmes des langues humaines a été démontrée chez des espèces aussi variées que les macaques, les chinchillas, les perruches, les pigeons, les gerbilles et les rats [40]. Le bonobo Kanzi parvient couramment à extraire les mots qu'il connaît d'un flot continu de parole [36]. Si, comme le prévoient certains modèles utilitaires coopératifs de l'évolution du langage, la parole bénéficiait à l'auditeur, la sélection naturelle aurait doté notre espèce de capacités acoustiques surdéveloppées sans rapport avec celles des autres primates et d'oreilles en forme de grands cornets orientables pour donner l'avantage à ceux qui sont capables de voler l'information à autrui ([28], p.351). Les capacités surdéveloppées de l'être humain, en ces matières, se situent plus dans la gorge que dans l'oreille.

Le système phonologique humain est un système combinatoire digital qui nous permet, au sein de chaque langue, d'assembler des sons pour constituer des morphèmes signifiants. Il s'agit d'un système particulièrement efficace qui permet d'émettre typiquement une quinzaine de signaux nouveaux par seconde. Encore une fois, il semble que ce système soit optimisé du côté du locuteur beaucoup plus que du côté de l'auditeur :

> "In fact, the temporal resolution capacity of the ear would not be good enough at normal speaking rates to

[1] Par exemple, près de 270 décès ont été causé par l'inhalation ou l'aspiration d'aliments en 1987 au Canada [24]. Le risque justifie que la manœuvre Heimlich fasse partie des techniques de base enseignées aux secouristes.

> segregate different phonemes and to perceive their proper order, if phonemes were consecutive bits of sound (Liberman & Mattingly 1985). Nature circumvents this limit imposed by the auditory system by packing the phonemes in such a way that each segment of sound conveys information about several phonemes." [1]

Certaines études de modélisations montrent que sous des hypothèses simples de contraste acoustique maximal et de déformation minimale, un simple tube doté de points de constriction évolue pour produire des voyelles et des consonnes proches de celles des langues humaines [7]. Ce résultat n'est pas obtenu par une coévolution de l'appareil émetteur et de 'appareil récepteur, mais seulement du premier, conformément à l'hypothèse selon laquelle la parole avantage avant tout le locuteur.

UN LEXIQUE PLÉTHORIQUE

La fonction biologique de l'appareil phonologique est indéniablement de permettre la communication non ambiguë de nombreux éléments lexicaux différents, et pour cela l'adoption d'un système combinatoire digital constitue une solution efficace [31]. On peut se demander, toutefois, quel avantage représente l'hypertrophie lexicale des langues humaines. Un humain adulte comprend des dizaines de milliers de mots de sa langue maternelle, sans compter les mots qu'il comprend dans les autres langues.[2] Quelle est la raison d'une telle inflation ? Pour certain auteurs, le langage a essentiellement une fonction utilitaire.

> "It is possible to imagine a superintelligent species whose isolated members cleverly negotiated their environment without communicating with one another, but what a waste! There is a fantastic payoff in trading hard-won knowledge with kin and friends, and language is obviously a major means of doing so" ([34], p. 367)

> "And, of course, the arrival of natural language would then have hugely facilitated both social co-operation and the acquisition knowledge. […] For its arrival would have made possible the detailed exchange of information, as well as the intricate but indefinitely flexible co-ordination of activity, which underlies much of the success of our species." ([8], p.231-232)

> "Language allowed our ancestors to share ideas and experiences, and to solve problems in parallel. The adaptive significance of human language is obvious. It pays to talk." [32]

Cependant, on voit mal comment la communication utilitaire pourrait justifier l'emploi de lexiques pléthoriques. Comme le remarque Robin Dunbar :

> […] most people would, at least until very recently, have supposed that [was language conveys] was related to information about hunting or the manufacture of tools. 'There were bison down at the lake yesterday when I was passing there' or 'If you want to make an arrowhead, you need to hit the flint nodule right here to strike off a suitable flake'. What is unsatisfactory about such claims is that (a) these kinds of technological activities take up a relatively small proportion of our time and (b) when we do engage in them, we actually rarely use language when doing so. Hunting is often best done in silence, and tool-making is best done by demonstration rather than instruction. ([22], p. 220)

De fait, la plupart des systèmes de communication utilitaire et de coordination se limitent à quelques dizaines de mots ou de signes, si l'on pense par exemple aux échanges liés au contrôle aérien ou à la plongée sous-marine. Les aspects matériels de l'environnement naturel d'*homo sapiens* étant relativement répétitifs, quelques mots et quelques gestes déictiques suffiraient à assurer l'essentiel de la coordination si telle était la fonction de la communication humaine. Or, le lexique de n'importe quelle langue contient une quantité de mots dont les sens sont quasi redondants, ce qui impose une charge considérable pour son apprentissage avec en retour un bien piètre gain, si l'efficacité se mesure à l'utilité des informations transmises.

> Why do we bother to learn so many rare words that have practically the same meanings as common words, if language evolved to be practical? ([28], p. 370)

La conclusion inévitable que l'on doit tirer de l'existence de lexiques démesurés est que le langage n'a pas pour fonction première de contribuer à la meilleure satisfaction des besoins humains.

Geoffrey Miller [28] suggère que la richesse du lexique résulte d'une compétition sexuelle amenant les individus à démontrer leur capacité à maîtriser de nombreux concepts. Cette théorie ne va pas sans poser plusieurs problèmes ; parmi les plus importants, citons le fait que les deux sexes participent également au langage dans notre espèce, alors que la théorie de Miller prédit nécessairement un dimorphisme sexuel important quant aux capacités de langage ; un autre problème est que cette théorie ne rend pas compte du contenu des prises de paroles spontanées, qui consiste essentiellement en narrations et en argumentation [16] qui intéressent les deux sexes, alors que la sélection sexuelle aurait dû le cantonner à l'expression de qualités viriles destinées à des auditoires féminins.

Le cadre théorique que nous avons proposé [16] [19] permet de justifier l'existence de lexiques de taille importante. Elle est liée à la qualité première affichée par le langage : montrer que l'on a su avant les autres. Prenons un exemple. Le 6 juillet 2005, 14h00 : certains de mes collègues sortent de leur bureau pour annoncer l'attribution inattendue des Jeux Olympiques de 2012 à la ville de Londres, quant tout le monde pensait que Paris serait choisie. Le comportement de ces collègues est incompréhensible s'il obéit à un réflexe communicationnel forgé pour des raisons utilitaires. Il est clair que tout le monde allait être au courant avant le soir. Si mes

[2] Le bilinguisme et le trilinguisme sont choses courantes dans les populations de chasseurs cueilleurs, étant donné l'aire géographique limitée des langues et les pratiques exogamiques systématiques.

collègues guettaient le résultat sur Internet et s'ils n'ont pas perdu une seconde pour l'annoncer, ce n'est pas pour augmenter le bien-être collectif, mais bien pour faire valoir leur capacité à savoir les premiers. Dans un tel schéma, ce n'est pas tant l'information elle-même qui importe, mais son originalité.

Si telle est la fonction première du langage : montrer que l'on détient des informations originales, on comprend pourquoi les lexiques ne peuvent pas se limiter à quelques dizaines de mots. Les événements originaux, qui permettent à ceux qui les annoncent de produire un effet sur leurs auditeurs, sont des événements inattendus, qui portent préférentiellement sur des faits rares [17]. Or, pour pouvoir désigner des états de fait rares, il est nécessaire de disposer d'un nombre suffisant de mots.

UNE MÉMOIRE DISPROPORTIONNÉE

Le lexique n'est pas le seul élément disproportionné de notre cognition. Les être humains stockent de nombreuses situations, toutes uniques, après une seule exposition. Cette capacité étonnante est, elle aussi, liée à la production de la parole.

L'observation des conversations quotidiennes montre que les individus consacrent la moitié du temps de parole, soit 10% de leur temps éveillé, à rapporter des faits passés qu'ils cherchent à présenter 'à propos' dans la conversation. L'originalité de tels récits réside dans leur caractère *inattendu*. Ainsi, passant dans une rue de Paris avec son amie, un locuteur montre une école et explique qu'il a été le témoin d'un incendie quelques années plus tôt, et qu'il pouvait voir les flammes sortir de la fenêtre de gauche. Ce comportement narratif qui emplit les conversations humaines est rendu possible par une particularité de notre espèce, la mémoire épisodique [41]. Certains animaux comme les écureuils ou les geais buissonniers (*Aphelocoma californica*) ont une mémoire spécialisée pour la rétention des caches de nourriture [10]. Les grands singes semblent avoir une mémoire de certains événements précis récents [37], mais la généralité, la précision et la taille de la mémoire épisodique humaine en font une capacité sans équivalent connu dans le monde vivant. Certains auteurs ont tenté de justifier cette capacité par son intérêt pratique [6], mais comme pour le lexique, le coût prohibitif d'un tel organe (la matière cérébrale qui stocke ces souvenirs consomme vingt fois plus d'énergie que les muscles [2]) ne saurait être compensé par la mémorisation d'épisodes entièrement instanciés qui ne contribuent que faiblement à l'efficacité de l'apprentissage. Le rôle de la mémoire épisodique, en revanche, prend tout son sens si les individus utilisent les épisodes qu'ils ont conservés en mémoire pour les évoquer dans les situations où ils peuvent apparaître inattendus.

Un événement est d'autant plus inattendu qu'il est plus simple à individualiser que prévu.[3] D'où l'importance de stocker des épisodes parfaitement instanciés. Prenons deux cas extrêmes : (1) le locuteur qui a vu une école en flammes en se rendant à son travail a trouvé la situation très inattendue au moment de l'événement, car elle était facile à individualiser au sein de toutes les classes de situations qui peuvent servir de référence ; (2) s'il mentionne simplement que telle école a brûlé dans le passé, son amie ne voit pas ce qui discerne cette école de tous les bâtiments devant lesquels le couple passe, ni ce qui discerne ce sinistre de tous les incendies qui se produisent chaque année dans Paris. L'effet conversationnel sera faible, et risque de provoquer un rejet du type *So what ?* [25]. Si le locuteur raconte qu'il a personnellement vu les flammes qui sortaient de cette fenêtre à gauche, l'effet sera intermédiaire entre (1) et (2) : en utilisant le témoignage instancié de son ami, l'interlocutrice peut plus facilement individualiser la situation et en apprécie le caractère inattendu.

Le comportement narratif, qui représente environ la moitié des prises de parole, requiert donc d'être capable de stocker et de décrire de manière non ambiguë de très nombreux épisodes. Là encore, la pression de sélection s'exerce sur les locuteurs. Sur le marché des alliances, où les liens sociaux se font et se défont, les individus capables de produire de l'inattendu sont davantage appréciés. L'évolution a favorisé ceux qui disposaient d'un vaste répertoire d'épisodes et des lexiques capables de les décrire de manière non ambiguë.

LE GESTE ET LA PAROLE

Certains auteurs défendent l'idée selon laquelle la communication humaine aurait été gestuelle avant d'être orale, et qu'il n'existe pas de continuité évidente entre les vocalisations des primates et le langage oral humain [11] [12] [3]. Le canal manuel-visuel présente de nombreux avantages en comparaison du canal vocal-auditif, ce qui justifie aux yeux de plusieurs auteurs l'idée selon laquelle son recrutement pour la communication référentielle ait été plus 'facile'. Leonard Talmy identifie un certain nombre d'avantages du canal manuel-visuel [38]. Citons :

- la présence d'une trentaine de paramètres pouvant varier indépendamment, contre 8 seulement pour le canal vocal-auditif
- un fort parallélisme
- des paramètres continus permettant une forte iconicité

Talmy tire argument du caractère fortement digital de la communication sur le canal vocal auditif pour expliquer la supériorité que celui-ci a finalement montrée pour la communication de concepts abstraits. Un argument similaire est proposé par Corballis selon qui l'accumulation des signifiés est incompatible avec un encodage analogique :

[3] D'un point de vue technique, l'inattendu U se mesure à la différence de complexité de description : $U = C_{exp} - C_{obs}$, où C_{exp} et C_{obs} désignent respectivement la complexité, au sens de Kolmogorov, attendue et observée [19].

> It would be difficult, for example, to make iconic signs that would distinguish ducks from drakes […] spoken words cannot be iconic representations of real-world objects or events. They can therefore be calibrated to minimize confusion between physically similar objects. ([12], p. 212).

L'argument est, cependant, discutable. L'exemple des langues des signes démontre que le canal manuel-visuel peut parfaitement servir de support à une transmission digitale. N'étant pas astreint à une iconicité stricte, il peut combiner les avantages des deux systèmes, tout en conservant son parallélisme et le nombre important des paramètres qui peuvent varier indépendamment. Dans ces conditions, le fait que la communication humaine soit spontanément orale et non gestuelle constitue un mystère.

La solution de cette énigme nous est fournie là encore par le constat que le langage a évolué dans l'intérêt du locuteur. Le problème qui se pose à tout locuteur, avant même de faire valoir la qualité de son message (notamment l'inattendu de la situation décrite dans le cas de la communication événementielle), est *d'attirer l'attention des auditeurs*. L'observation des conversations quotidiennes montre qu'elles mettent en scène une compétition, non du côté des auditeurs, mais des locuteurs.

> Watch any group of people conversing, and you will see the exact opposite of the behaviour predicted by the kinship and reciprocity theories of language. People compete to say things. They strive to be heard. ([28], p. 350)

Dans un tel contexte compétitif, attirer l'attention des autres devient crucial. Sur ce registre, le canal vocal-auditif possède un avantage décisif sur le canal manuel-visuel, car l'attention auditive des individus est beaucoup plus facile à *forcer* que leur attention visuelle. Il suffit d'observer un groupe de malentendants en train de signer pour s'en persuader. Si l'on suit cet argument, on doit admettre que le langage a eu une composante orale dès ses débuts, dès le stade du geste référentiel déictique. L'argument n'est pas incompatible avec une coévolution de la parole et du geste spontané, au contraire. En revanche, en conférant le primat à la modalité vocale, il exclut les scénarios de type "échafaudage" qui imaginent que l'émergence du langage a dû s'appuyer sur un état antérieur purement gestuel de la communication humaine.

CONCLUSION

Nous avons évoqué dans ce qui précède plusieurs conséquences du fait que le langage a évolué au bénéfice des locuteurs et non, conformément à une idée répandue, au bénéfice des auditeurs. L'activité langagière humaine constitue l'une des *arènes* où se jouent l'établissement et le maintien des liens de solidarité. Compte tenu de leur structure sociale particulière, caractérisée par des coalitions de taille importante, les humains utilisent plusieurs critères pour le choix de leurs alliés. Parmi ces critères, la capacité informationnelle, par laquelle les individus connaissent l'état de leur environnement physique et social, joue un rôle important. Cette préférence pour les individus bien informés engendre une compétition dans laquelle les locuteurs rivalisent pour afficher leur compétence informationnelle, ce qu'ils font notamment en rapportant tout fait qui peut apparaître inattendu.

Les conséquences de cette compétition sont multiples et en conformité avec ce que nous observons du langage tel qu'il est pratiqué. Nous avons mentionné le caractère non utilitaire de la plupart des prises de parole, le fait que l'appareil phonatoire a évolué bien davantage que l'appareil auditif, le fait que les lexiques atteignent une taille disproportionnée, le fait que les humains maintiennent en mémoire une grande quantité d'épisodes instanciés, et enfin le fait que la communication humaine passe par la parole, alors que tout prédestinait la modalité gestuelle dans ce rôle si le critère premier était l'efficacité.

Le scénario qui voit dans le langage un moyen d'affichage des capacités informationnelles est riche d'autres prédictions. Par exemple, il explique comment l'argumentation a pu naître comme un moyen pour les auditeurs de lutter contre le mensonge, qui permet de produire de l'inattendu à moindre frais lorsque les faits rapportés ne sont pas vérifiables [16].

Le point crucial qui reste à éclaircir réside dans le changement d'organisation social qu'a connu la lignée humaine. L'hypothèse concernant l'importance pour les humains d'afficher leurs capacités informationnelles par le langage repose sur le fait que la connaissance de l'environnement physique et social est essentielle pour la prise de décision collective au sein d'une coalition. Ceci n'a de sens que si les coalitions sont de taille significative, quelque cinq ou dix individus. Chez les chimpanzés mâles, les coalitions n'excèdent pas trois individus [13]. La qualité la plus recherchée est la force physique, et c'est aussi celle qui est préférentiellement affichée. Pour une raison qui reste inconnue, notre lignée a connu une bifurcation, avec l'émergence de coalitions plus vastes. C'est ce changement qui a conféré son importance à la capacité informationnelle, et c'est à ce changement que nous devons de parler.

BIBLIOGRAPHIE

[1] O. Aaltonen & E. Uusipaikka. "Why speaking is so easy? - Because talking is like walking with a mouth". In: M. Suominen, A. Arppe & et al. (Eds), *A man of measure: Festschrift in honour of Fred Karlsson*. A special supplement to SKY Journal of Linguistics vol. 19:111-118, 2006.

[2] L. C. Aiello. "Brains and guts in human evolution: The Expensive Tissue Hypothesis". *Brazilian Journal of Genetics* 20(1):141-148, 1997.

[3] M. A. Arbib. "Grounding the mirror system hypothesis for the evolution of the language-ready brain". In: A. Cangelosi & D. Parisi (Eds),

Simulating the evolution of language. Springer Verlag, London, pages 229-254, 2001.

[4] R. Barone. *Anatomie comparée des mammifères domestiques*. Vigot, Paris, 1976.

[5] I. Brinck & P. Gärdenfors. "Co-operation and communication in apes and humans". *Mind and Language* 18(5):484-501, 2003.

[6] R. Brown & J. Kulik. "Flashbulb memories". *Cognition* 5:73-99, 1977.

[7] R. Carré. "From an acoustic tube to speech production". *Speech communication* 42:227-240, 2004.

[8] P. Carruthers. *Language, Thought and Consciousness*. Cambridge University Press, Cambridge, MA, 1996.

[9] N. Chomsky. *Réflexions sur le langage*. Flammarion, Paris, 1975 (ed. 1981).

[10] N. S. Clayton & A. Dickinson. "Episodic-like memory during cache recovery by scrub jays". *Nature* 395:272-274, 1998.

[11] M. C. Corballis. "Did language evolve from manual gestures?". In: A. Wray (Ed), *The transition to language*. Oxford University Press, Oxford, UK, pages 161-179, 2002.

[12] M. C. Corballis. "From hand to mouth: The gestural origins of language". In: M. H. Christiansen & S. Kirby (Eds), *Language Evolution*. Oxford University Press, Oxford, pages 201-218, 2003.

[13] F. B. M. de Waal. *Chimpanzee politics: power and sex among apes*. The John Hopkins Univ. Press, Baltimore, 1982 (ed. 1989).

[14] T. W. Deacon. *The symbolic species*. W.W. Norton & Co., New York, NY, 1997.

[15] J-L. Dessalles. "Coalition factor in the evolution of non-kin altruism". *Advances in Complex Systems* 2(2):143-172, 1999.

[16] J-L. Dessalles. *Aux origines du langage : Une histoire naturelle de la parole*. Hermès-sciences, Paris, 2000.

[17] J-L. Dessalles. "Vers une modélisation de l'intérêt". In: A. Herzig, Y. Lespérance & A-I. Mouaddib (Eds), *Actes des troisièmes journées francophones 'Modèles formels de l'interaction' (MFI-05)*. Cépaduès Editions, Toulouse, 113-122, 2005.

[18] J-L. Dessalles. "A structural model of intuitive probability". In: D. Fum, F. Del Missier & A. Stocco (Eds), *Proceedings of the seventh International Conference on Cognitive Modeling*. Edizioni Goliardiche, Trieste, IT, pages 86-91, 2006.

[19] J-L. Dessalles. "Generalised signalling: a possible solution to the paradox of language". In: A. Cangelosi, A. D. M. Smith & K. Smith (Eds), *The evolution of language*. World Scientific, Singapore, pages 75-82, 2006.

[20] R. I. M. Dunbar, N. Duncan & D. Nettle. "Size and structure of freely forming conversational groups". *Human nature* 6(1):67-78, 1995.

[21] R. I. M. Dunbar. *Grooming, gossip, and the evolution of language*. Harvard University Press, Cambridge, MA, 1996.

[22] R. I. M. Dunbar. "The origin and subsequent evolution of language". In: M. H. Christiansen & S. Kirby (Eds), *Language Evolution*. Oxford University Press, Oxford, UK, pages 219-234, 2003

[23] H. Gintis, E. A. Smith & S. Bowles. "Costly Signaling and Cooperation". *Journal of Theoretical Biology* 213:103-119, 2001.

[24] R. B. Goldbloom. *The Canadian guide to clinical preventive health care*. Canada Communication Group – Publishing, Ottawa, CA, 1994.

[25] W. Labov. "Some further steps in narrative analysis". *Journal of Narrative and Life History* 7(1-4):395-415, 1997.

[26] A. M. Liberman & I. G. Mattingly. "The motor theory of speech perception revised". *Cognition* 21(1):1-36, 1985.

[27] J. Maynard Smith. *The Theory of Evolution*. Penguin Books, New York, NY, 1958 (ed. 1975).

[28] G. F. Miller. *The mating mind*. Doubleday, New York, NY, 2000.

[29] F. J. Newmeyer. "What can the field of linguistics tell us about the origins of language?". In: M. H. Christiansen & S. Kirby (Eds), *Language Evolution*. Oxford University Press, Oxford, UK, pages 58-76, 2003.

[30] W. Noble & I. Davidson. *Human evolution, language and mind*. Cambridge University Press, Cambridge, MA, 1996.

[31] M. A. Nowak, D. C. Krakauer & A. Dress. "An error limit for the evolution of language". *Proceedings of the Royal Society of London* B266:2131-2136, 1999.

[32] M. A. Nowak & N. L. Komarova. "Towards an evolutionary theory of language". *Trends in cognitive sciences* 5(7):288-295, 2001.

[33] S. Pinker & P. Bloom. "Natural language and natural selection". *Behavioral and Brain Sciences* 13(4):707-784, 1990.

[34] S. Pinker. *The language instinct*. Harper Perennial, New York, NY 1994 (ed. 1995).

[35] S. Pinker. "Language as an adaptation to the cognitive niche". In: M. H. Christiansen & S. Kirby (Eds), *Language Evolution*. Oxford University Press, Oxford, UK, pages 16-37, 2003.

[36] E. S. Savage-Rumbaugh & R. Lewin. *Kanzi: the ape at the brink of the human mind*. John Wiley & Sons, New York, NY, 1994.

[37] B. L. Schwartz, M. L. Hoffman & S. Evans. "Episodic-like memory in a gorilla: A review and new findings". *Learning and Motivation* 36:226-244, 2005.

[38] L. Talmy. "Recombinance in the evolution of language". In: J. E. Cihlar, D. Kaiser & Irene Kimbara (Eds), *Proceedings of the 39th Annual Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago, IL, 2004.

[39] M. Tomasello. *The cultural origins of human cognition*. Harvard university press, Cambridge, MA, 1999.

[40] J. M. Toro, J. B. Trobalon & N. Sebastián-Gallés. "Effects of backward speech and speaker variability in language discrimination by rats". *Journal of Experimental Psychology: Annimal Behavior Processes* 31(1):95-100, 2005.

[41] E. Tulving. *Elements of episodic memory*. Oxford University Press, New York, NY, 1983.

[42] I. Ulbaek. "The origin of language and cognition". In: J. R. Hurford, M. Studdert-Kennedy & C. Knight (Eds), *Approaches to the evolution of language: social and cognitive bases*. Cambridge University Press, pages 30-43, Cambridge, MA, 1998.

[43] G. C. Williams. *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton University Press, Princeton, NJ, 1966 (ed. 1996).

[44] A. Zahavi & A. Zahavi. *The handicap principle*. Oxford University Press, New York, NY, 1997.

Session III

Reconnaissance de la parole

Lundi 12 juin 2006 - 15h30 16h30

Expériences de transcription automatique d'une langue rare

Thomas Pellegrini and Lori Lamel

LIMSI-CNRS, BP133
91403 Orsay cedex, FRANCE
{thomas.pellegrini, lamel}@limsi.fr

ABSTRACT

This work investigates automatic transcription of rare languages, where rare means that there are limited resources available in electronic form. In particular, some experiments on word decompounding for Amharic, as a means of compensating for the lack of textual data are described. A corpus-based decompounding algorithm has been applied to a 4.6M word corpus. Compounding in Amharic was found to result from the addition of prefixes and suffixes. Using seven frequent affixes reduces the out of vocabulary rate from 7.0% to 4.8% and total number of lexemes from 133k to 119k. Preliminary attempts at recombining the morphemes into words results in a slight decrease in word error rate relative to that obtained with a full word representation.

1. INTRODUCTION

Les termes utilisés dans la littérature pour désigner les langues rares sont variés, que ce soit en français : langues rares, langues peu-dotées, langues peu-informatisées (langues pi), langues minoritaires, ou que ce soit en anglais : under-represented languages, under-resourced languages, less widely available languages ou encore minority languages. Cette diversité de vocabulaire reflète le caractère subjectif de la notion de rareté pour les langues, du point de vue du traitement automatique. Dans le rapport final du projet européen INTERA [5] qui visait la création de nouveaux corpus de ressources multilingues pour des langues européennes, les langues visées furent des langues "moins accessibles du point de vue numérique". Cette définition relève d'une comparaison implicite avec les langues dominantes, et parait plus précise que le simple adjectif "rare" employé dans un souci de concision dans cet article. L'amharique, langue officielle de l'Ethiopie, fait l'objet de recherches récentes en traitement automatique telles que le classement thématique de textes [2], les outils d'analyse morphologique [3] et la transcription automatique [6],[9]. Pour cette langue, l'audio n'est pas un facteur limitant puisque de nombreuses radios diffusent quotidiennement leurs émissions sur Internet [1]. En revanche il s'agit d'une langue rare en raison de la faible quantité de textes disponibles sous forme électronique. Le corpus de textes utilisé dans cette étude ne dépasse pas 5 millions de mots, chiffre à comparer au milliard de mots utilisés pour les systèmes de transcription d'anglais américain et du français du LIMSI.

Dans un premier temps, les propriétés lexicales du corpus de textes et l'élaboration d'un lexique de prononciation seront décrites, les résultats d'un système standard de transcription automatique seront présentés. Dans un deuxième temps les premiers résultats d'expériences de décomposition des mots (par séparation d'affixes) seront discutés.

2. PRÉSENTATION DE L'AMHARIQUE

L'amharique est parlé par environ 14 millions de locuteurs. Bien que faisant partie des langues sémitiques comme l'arabe et l'hébreu, l'écriture se fait de gauche à droite avec un syllabaire spécifique dérivé de la langue classique éthiopienne, le ge'ez. L'amharique possède 34 symboles de base dont 85% représentent une séquence CV (C pour consonne, V pour voyelle), les autres symboles représentent une séquence CwV où w est une semi-consonne. Un dernier symbole représente le son complexe /ts/. Cette langue possède sept voyelles au total, schwa inclus, appelées les sept ordres : /ɛ/, /u/, /i/, /a/, /e/, /ə/ et /o/.

En ce qui concerne l'écrit, des problèmes de normalisation de l'orthographe amharique sont exposés dans [10], où trois niveaux de langue sont distingués :

- l'amharique canonique, réservé aux érudits (une orthographe unique par mot),
- l'amharique commun, celui des journaux, de la littérature (formes homophones),
- l'amharique quotidien, pas de jugement porté sur l'orthographe des mots.

L'orthographe des mots amhariques utilisés au quotidien est très libre, le nombre de formes écrites différentes pour un même mot peut être très grand. Des exemples d'orthographe ambigue rencontrés seront donnés ci-après.

3. LES RESSOURCES AUDIO ET TEXTES

Pour élaborer le système de transcription et réaliser les études décrites dans cet article, deux types de données ont été utilisés : un corpus d'émissions de radio transcrites et un corpus de textes issus de sites Web de journaux en ligne. Le corpus audio contient 37 heures d'émissions de type journal provenant de 2 sources : radio Deutsche Welle (25h) et radio Medhin (12h) enregistrées de janvier 2003 à février 2004. Ces données ont été transcrites manuellement par des locuteurs ethiopiens. Pour tester et développer le système de transcription automatique, deux heures de données audio transcrites ont été sélectionnées au sein du corpus. Il s'agit des fichiers audio parmi les plus récents du corpus, les thèmes abordés dans ces données pouvant être nouveaux par rapport à ceux des données d'apprentissage. Le tableau 1 résume les caractéristiques des données

[1] Deutsche Welle et Radio Medhin par exemple

audio : le nombre d'heures par source, le nombre de locuteurs et le nombre de mots pour le corpus d'apprentissage et pour le corpus de développement. Nous avons retenu un plus grand nombre d'heures provenant de Deutsche Welle que de radio Medhin, car les émissions de cette radio ont une plus grande diversité de locuteurs.

TAB. 1: Nombre d'heures, de locuteurs et de mots pour les deux sources audio

source	app	dev
Deutsche welle	24h06	1h20
radio Medhin	11h08	0h37
# locuteurs	200	15
# mots	232.6k	14.1k

Les données textuelles autres que les transcriptions manuelles proviennent de 3 sources : Ethiozena (archives de 1988 a 1996 et textes récents), Deutsche Welle (textes récents) et Ethiopian Reporter (textes récents). Au total pour ces trois sources nous disposons de 4.6 millions de mots. Les textes des transcriptions du corpus audio sont également utilisées et totalisent 246.7k mots.

4. PROPRIÉTÉS LEXICALES

A l'instar de l'arabe et de l'hébreu un grand nombre de mots, en particulier les verbes, se forment à partir de racines de trois lettres (racines tri-consonantiques) auxquelles s'ajoutent des schèmes (les voyelles) qui précisent le sens des mots ainsi formés [3]. Viennent s'ajouter des marques (de conjugaison, de pronoms personnels, de pluriel, etc…) qui donnent de nombreuses formes dérivées à partir d'une même racine.

La figure 1 montre le nombre de mots distincts en fonction de leur taille en nombre de phonèmes pour les transcriptions manuelles (courbe en trait plein) et pour les textes Web (courbe en pointillé). La taille de mots la plus fréquente est relativement grande puisqu'elle est de dix phonèmes soit cinq syllabes (cinq symboles amhariques), ce qui pourrait justifier la démarche de chercher des affixes pour décomposer les mots longs.

La figure 2 donne la taille moyenne des mots en fonction du rang de fréquence des mots distincts des textes Web. La courbe montre que les mots les plus fréquents sont les plus courts avec une taille entre deux et trois syllabes, ce qui est un comportement tout à fait naturel, observé pour une majorité de langues. Néanmoins le corpus de textes étant très petit, le nombre de mots distincts peu fréquents est grand. Pour un rang de fréquence compris entre 1 et 500, il y a un seul mot par rang. A partir de 500, ce nombre augmente et atteint, pour les mots les moins fréquents (1300^e rang, 3 occurences dans le corpus), une valeur supérieure à 26k mots. Pour les textes Web, seuls les mots apparaissant au moins trois fois sont gardés pour se débarrasser le plus possible des mots "parasites" (chiffres accolés aux mots, caractères non-identifiés, etc…)

Les dix premiers rangs de fréquence couvrent un peu plus de 5% des 4.6M de mots des textes Web (dix mots distincts totalisant 245.8k occurences). Les dix derniers rangs de fréquence couvrent 10% (80k mots distincts totalisant 423.9k occurences). Les expériences de décomposition sur les mots longs (moins fréquents mais nombreux) peuvent donc s'avérer intéressantes pour les performances du système de transcription, à condition que les racines des mots dont les affixes ont été séparés soient des mots déjà présents avant décomposition.

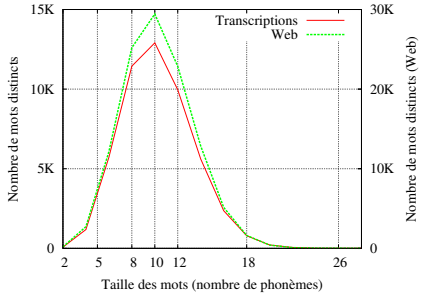


FIG. 1: Distribution des mots distincts des transcriptions et des textes Web en fonction de leur taille (en nombre de phonèmes)

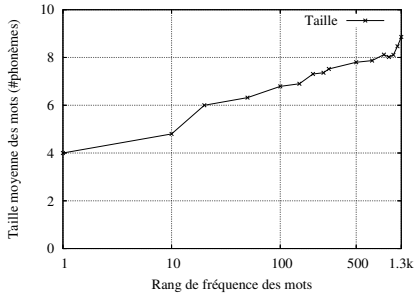


FIG. 2: Taille moyenne des mots distincts des textes Web (112k mots) en fonction de leur rang de fréquence

Le taux de mots hors vocabulaire (MHV) influence grandement les performances des systèmes de transcription (1 mot hors vocabulaire génère entre 1.5 et 2 erreurs). Les systèmes actuels travaillent avec des lexiques fixes qui doivent optimiser la couverture lexicale. Avec un lexique de 133k mots réalisé avec tous les mots des transcriptions et les mots des textes Web apparaissant au moins trois fois, le taux de MHV atteint 7.0% sur le corpus de développement de 14.1k mots. A titre de comparaison, le taux de MHV pour le système de transcription de l'anglais américain est d'environ 0.5% avec un lexique de 65k mots [4]. Tenter de le diminuer en décomposant les mots longs est une autre motivation importante pour cette étude.

5. LE LEXIQUE DE PRONONCIATIONS

Le jeu de phonèmes utilisé dans ces expériences comporte 33 phones (voyelles comprises) avec trois phones supplémentaires (pour les silences, les respirations et les hésitations, annotés dans les transcriptions). Les suites Cw sont modélisées avec deux unités distinctes, une pour C, une autre pour w. Si aux symboles amhariques qui codent les suites Cw avaient été associés un unique caractère de translitération, le nombre de phones aurait été plus grand. Dans le système de reconnaissance décrit dans [9] par exemple, le jeu de phones en compte 38 soit 5 consonnes de plus.

Les 240 symboles amhariques ont été translitérés en utilisant le jeu de phones mentionné ci-dessus. Voici un exemple de phrase tirée d'une transcription manuelle, il s'agit d'une phrase d'introduction des émissions d'information de la radio Medhin (? pour le coup de glotte, x pour le schwa) :

የኢትዮጵያ መድሀን ድምፅ ራዲዮ
jE ?itxjoPxja mEdxhxnx dxmxtsx radijo

Comme les caractères amhariques sont translitérés avec un jeu de caractères représentant les phonèmes eux-mêmes, un premier dictionnaire de prononciation constitué de la simple liste des mots a été utilisé pour réaliser les premiers alignements. Ces alignements ont permis de voir que les schwas ne sont pas toujours prononcés et un deuxième dictionnaire avec tous les schwas optionnels a été utilisé. Dans l'article [8], des expériences de détection des variantes de prononciation des voyelles amhariques à l'aide d'alignements sont décrites. Des alignements de syllabes sont réalisés avec un lexique permettant la substitution et l'élision de toutes les voyelles. Globalement, plus de 60% des schwas sont éliminés. Le lexique utilisé dans cette étude est un lexique de 133k mots avec uniquement les schwas optionnels. Le tableau 2 donne un exemple de trois mots de ce lexique, avec leur rang de fréquence et leur nombre d'occurences au sein des transcriptions manuelles. Le schwa est mis entre accolades pour signifier son caractère optionnel dans les formes phonémiques du lexique. Sur les 3k occurences de "nEwx", 2.5k ont été alignées avec la prononciation sans schwa.

TAB. 2: Exemples de mots du lexique avec leur prononciation, rang de fréquence et nombre d'occurences

lexème	forme phonémique	rang	#occ.
nEwx	nEw{x}	1	3044
mEto	mEto	7	803
jEdimokxrasi	jEdimok{x}rasi	236	47

6. RÉSULTATS AVEC UN SYSTÈME STANDARD

Un système de transcription standard basé sur [4], en deux passes avec une adaptation non-supervisée des modèles acoustiques après la première passe a été construit avec la boîte à outils STK du LIMSI. Le modèle de langage est issu de l'interpolation de deux modèles avec lissage de Kneser-Ney, le premier est un modèle appris uniquement sur les transcriptions et le second sur les textes Web. La perplexité mesurée sur le corpus de développement est relativement élevée : $px = 372$ avec un taux de MHV de 7.0%. Les modèles acoustiques utilisés sont des triphones à états liés (avec 32 gaussiennes par état, 3 états par modèle) dépendants du contexte et de la position inter-mot, i.e. différents modèles sont utilisés pour des phones à l'intérieur des mots et pour des phones en frontière de mots. Au total, 10.7k contextes sont modélisés par 8.5k états. Les contextes peu observés sont regroupés en regardant en premier lieu s'ils ont un contexte droit en commun (co-articulation régressive), sinon un contexte gauche en commun (co-articulation progressive) et enfin ceux pour lesquels aucun contexte en commun n'a été trouvé sont regroupés en modèles indépendants du contexte.

Le taux d'erreur sur les mots évalué à l'aide de l'outil sclite de NIST sur le corpus de développement est de 25.9%. Le tableau 3 précise les pourcentages des trois types d'erreur (E pour élision, S pour substitution et I pour insertion). L'équilibre entre les taux d'élisions et d'insertion est obtenu par une phase de réglage des paramètres du décodeur (le poids du modèle de langage, les pénalités sur le nombre de mots, le nombre de silences entre autres).

Le tableau 4 montre des exemples d'erreurs fréquentes en donnant la référence puis l'hypothèse erronnée avec les types d'erreur correspondants. Dans le premier cas le système a reconnu un mot composé plutôt que deux mots plus petits ce qui a donné une élision et une substitution. Le deuxième exemple est l'inverse du premier. Le système a opté pour deux mots plus petits plutôt qu'un seul mot ce qui a donné une insertion et une substitution. Enfin le troisième exemple montre un problème a priori de normalisation orthographique, les deux mots ne différant que d'une voyelle, néanmoins les deux formes sont présentes dans les textes.

TAB. 3: Répartition des erreurs. Types d'erreur : E élision, S substitution et I insertion

S	E	I	Total
20.9%	2.6%	2.3%	25.9%

TAB. 4: Exemples d'erreurs fréquentes. REF : référence, HYP : hypothèse. Types d'erreur : E élision, S substitution et I insertion

	exemple	type d'erreur
REF	bEjE KEnu	
HYP	bEjEKEnu	E S
REF	?xnxdEdxlx	
HYP	?xnxdE dxlx	I S
REF	kE ?**e**konomiwx	
HYP	kE ?**i**konomiwx	S

7. SÉPARATION D'AFFIXES : PREMIERS RÉSULTATS

La taille du lexique et le fort taux de MHV peuvent être diminués en séparant des affixes. Des expériences de décomposition de mots ont été réalisées pour des langues où le procédé de composition des mots est très important, pour l'allemand par exemple [1]. En amharique la composition se limite à des morphèmes grammaticaux de type articles possessifs, démonstratifs, pronoms, prépositions et postpositions.

7.1. Le choix des affixes

Détecter les affixes automatiquement a l'avantage de ne pas utiliser de connaissances linguistiques spécifiques à la langue cible et rend la méthode portable à d'autres langues facilement. L'algorithme de Harris [7] est un algorithme de détection des frontières de morphèmes indépendant de la langue, nécessitant simplement un corpus de mots de la langue cible. Il exploite le fait qu'un début de mot de k caractères a naturellement peu de caractères successeurs distincts possibles pour former des mots qui existent dans la langue traitée pour k suffisamment grand. Au rang k+1 ce nombre réduira davantage. Si ce nombre augmente subitement pour un début de mot de k caractères alors ce début de mot est un morphème candidat, pouvant se composer avec d'autres morphèmes commençant par des lettres distinctes variées. Ainsi l'algorithme compte le nombre de caractères successeurs distincts possibles pour tous les débuts de mots de taille k et propose des frontières de morphèmes pour ces mots lorsqu'un maxima local est trouvé. Il ne s'agit pas de réaliser une analyse morphologique mais de dégager quelques affixes potentiels les plus fréquents. Séparer ces affixes des mots du lexique permet de réduire le nombre de lexèmes et d'augmenter la représentation de certains n-grammes peu observés [1].

Une liste de sept affixes (cinq préfixes et deux suffixes) a été retenue pour les premières expériences rapportées ici.

Les sept affixes sont les plus fréquents parmi ceux détectés par l'algorithme. Les affixes sont séparés des mots dont la taille après séparation est d'au moins deux syllabes. Un signe "+" est accolé aux affixes pour pouvoir recombiner les mots par la suite. Le tableau suivant donne la liste des affixes :

TAB. 5: Préfixes et affixes retenus

préfixes (5)	suffixes (2)
?xnxdE+	+CEwx
?xnxda+	+mx
?xnxdi+	
?xnxdx+	
jE+	

Le nombre de mots du lexique est réduit de plus de 11%, de 133k à 119k mots. Le taux de MHV diminue de 7.0% à 4.8% soit une réduction absolue de 2.2%.

7.2. Résultats

De nouveaux modèles acoustiques ont été appris pour la représentation avec affixes séparés et un nouveau modèle de langage a été généré. Le sytème de reconnaissance est le même que celui qui a servi pour la représentation en mots entiers, seuls les modèles acoustiques, le modèle de langage et le lexique de prononciation diffèrent. Les résultats suivants ont été obtenus avec 10.6k modèles acoustiques (8.7k états).

Le tableau 6 donne les taux d'erreur pour la représentation avec les affixes séparés et après recombinaison des mots. Le taux obtenu avec affixes séparés est nettement inférieur car le nombre de mots est plus grand avec cette representation (17.3k mots), les affixes étant très bien reconnus. En recombinant les affixes (grâce au signe "+" accolé), le taux d'erreur augmente. Un gain absolu de 0.8% est néanmoins observé par rapport au taux d'erreur de 25.9% avec le système appris sur les mots entiers (appelé S_{mots} par la suite).

TAB. 6: Taux d'erreur sur les mots avant et après recombinaison des affixes

représentation	taux d'erreur
affixes séparés	21.6%
mots recomposés	25.1%

Le tableau 7 donne un exemple où le système avec affixes séparés a été meilleur. Les phrases de référence, en gras dans le tableau, ont respectivement trois mots pour la représentation en mots entiers (S_{mots}) et quatre mots pour la représentation avec affixes séparés ($S_{affixes}$). Le système S_{mots} n'a pas bien reconnu la phrase, la sortie obtenue ayant deux mots au lieu de trois. En revanche le système $S_{affixes}$ a correctement reconnu la phrase de référence. Le tableau donne, pour chaque système, la log-vraisemblance (log-v) pour la phrase correcte (phrase de référence) et pour la phrase erronée résultat du décodage par le système S_{mots}. Pour $S_{affixes}$, la phrase erronée est la phrase erronée de S_{mots} après séparation de l'affixe.

La vraisemblance, qui est la probabilité des suites de mots, est utilisée par le décodeur pour sélectionner la meilleure hypothèse. Les mots "lajx" et "jE+" sont parmi les mots les plus fréquents des textes avec affixes séparés, alors que le mot " ?iraKxlajx"est beaucoup moins fréquent, ce qui favorise la phrase correcte, obtenue par $S_{affixes}$, qui contient "lajx"et "jE+". Pour le système S_{mots}, la forte probabilité de l'unigramme "lajx" ne suffit pas à favoriser

TAB. 7: Exemple de phrase correctement reconnue par $S_{affixes}$ mais erronée pour S_{mots}, comparaison des log-vraisemblances. Les lignes S_{mots} et $S_{affixes}$ donnent les sorties respectives des systèmes

sytème	phrase		log-v
S_{mots}	?iraKxlajx	jESxgxgxrx	-9.5551
	?iraKx **lajx**	**jESxgxgxrx**	-9.6559
$S_{affixes}$	**?iraKx** **lajx**	**jE+ Sxgxgxrx**	-9.2613
	?iraKxlajx	jE+ Sxgxgxrx	-10.1367

la phrase de référence. Le mot "jESxgxgxrx" étant rare, la séparation de l'affixe "jE+" a été bénéfique.

8. PERSPECTIVES

Dans cet article, un système de transcription de l'amharique a été présenté et a servi de référence pour évaluer un deuxième système construit avec une représentation des mots différente, avec séparation d'affixes. La séparation d'un très petit nombre d'affixes a permis de réduire la taille du lexique de plus de 11%, le taux de MHV de 2.2% absolu et d'observer un gain de 0.8% absolu sur le taux d'erreur sur les mots. De nouvelles expériences avec un nombre d'affixes plus grand seraient très intéressantes à mener, la question de la sélection des affixes étant problématique (leur nombre, leur ressemblance phonémique). Le découpage des mots augmente la représentation de certains ngrammes peu observés et semble être une méthode prometteuse pour aborder le problème de la transcription automatique de langues rares.

RÉFÉRENCES

[1] M. Adda-Decker. A corpus-based decompounding algorithm for German lexical modeling in LVCSR. In *Proc. Eurospeech*, Geneva, 2003.

[2] S. Eyassu and B.Gamback. Classifying Amharic news texts uing self-organizing Maps. In *ACL05 Workshop on computational Approaches to Semitic Languages*, Ann Arbor, Michigan, 2005.

[3] S. Fissaha and J. Haller. Amharic verb lexicon in the context of machine translation. In *TALN 2003*, Batz-sur-Mer, 2003.

[4] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37 :89–108, 2002.

[5] M. Gavrilidou, V. Giouli, E. Desipri, M. Monachini, and C. Soria. Report on the model of LRs production.

[6] B. Gamback H. Seid. A speaker independent continuous speech recognizer for Amharic. In *Proc. Interspeech*, Lisboa, 2005.

[7] Z. Harris. From Phoneme to Morpheme. In *Language 31*, pages 190–222, 1996.

[8] T. Pellegrini and L. Lamel. Experimental detection of vowel pronunciation variants in Amharic. In *Proc. LREC*, Genoa, 2006.

[9] W. Menzel S.T. Abate and B. Tafila. An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition. In *Proc. Interspeech*, Lisboa, 2005.

[10] D. Yacob. Application of the Double Metaphone Algorithm to Amharic Orthography. In *International Conference of Ethiopian Studies XV*, 2003.

Reconnaissance automatique de la parole en langue somalienne

Abdillahi Nimaan [1,2], Pascal Nocera[1] et Jean-François Bonastre[1]

[1] Laboratoire Informatique d'Avignon (Université d'Avignon et des pays du Vaucluse)
BP 1228 84911 Avignon Cedex 9, France
[2] Institut des Sciences et des Nouvelles Technologies (Centre d'Études et des Recherches de Djibouti)
BP. 486 Djibouti, Djibouti
{nimaan.abdillahi, pascal.nocera, jean-françois.bonastre}@univ-avignon.fr
http ://www.lia.univ-avignon.fr

ABSTRACT

Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. Automatic transciption and indexing tools seem potential solution to preserve it. This paper presents the first results of automatic speech recognition (ASR) of Djibouti languages in order to index the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected and the first ASR results of this language. Using the specificities of the Somali language, (words are composed of a concatenation of sub-words called "roots" in this paper), we improve the obtained results. We also discuss future ways of research like roots indexing of audio archives.

1. INTRODUCTION

Le patrimoine scientifique, culturel et historique des pays africains se transmet oralement de générations en générations. Ce savoir ancestral, accumulé durant des siècles est menacé de disparition, du fait du processus de mondialisation, de la transformation sociale ainsi que du manque de moyen de sauvegarde. De nombreuses organisations nationales et internationales [15] oeuvrent pour endiguer ce phénomène. Aujourd'hui, la plupart des pays concernés disposent d'importantes bases de données audio, archivées, le plus souvent, par les stations radio locales depuis plusieurs décennies. Ces pays sont confrontés à deux questions : sauvegarder ce patrimoine par un programme de numérisation et le rendre plus accessible. Concernant le premier point, les techniques sont bien connues, et la numérisation, en cours dans de nombreux pays, n'est qu'un problème d'ordre logistique. Le second point est plus délicat, car l'exploitation de bases de données audio de grandes tailles [1] nécessite des traitements informatiques de haut niveau pour toutes les langues des pays concernés, tels que des outils de transcription et d'indexation automatiques. Ce papier présente les prémices du traitement automatique du patrimoine culturel audio de la république de Djibouti. Dans un premier temps, les langues djiboutiennes et les différents corpus constitués pour cette étude sont présentés. Nous décrivons ensuite les expériences de reconnaissance de la parole somalienne effectuées sur les mots et sur les racines. Finalement, nous tirons les conclusions de ces travaux, et énonçons les futurs axes de recherche.

2. LANGUES DJIBOUTIENNES

Quatre langues sont parlées à Djibouti, le français et l'arabe sont officiels, l'afar et le somalien sont autochtones et largement utilisés. Nos travaux actuels portent uniquement sur la langue somalienne qui concerne la moitié des archives audio ciblées. Cette langue est parlée dans plusieurs autres pays de l'Afrique de l'est par une population estimée entre 12 et 15 millions [2]. Elle est répertoriée dans la sous famille couchitique des langues afro-asiatique dans la classification internationale SIL [3]. La variante somali-somali, communément appelée langue somalienne, et parlée à Djibouti, est plus précisement visée dans nos recherches. Son système phonétique est composé de 22 consonnes et de 20 voyelles (5 longues et 5 courtes avec $\pm$ ATR [4]) [14]. La table 1 présente la structure phonétique des consonnes. C'est également une langue tonale avec deux ou trois tons différents [6], [7], [13]. Sa forme graphique est relativement jeune, puisqu'elle n'est écrite que depuis 1972 en caractères latins. Il n'existe donc aucun document écrit antérieur à cette date. La transcription d'un mot est directement issue de sa réalisation phonétique (chaque phonème est représenté par une lettre).

3. CONSTITUTION DES CORPUS

3.1. Corpus textuel

La reconnaissance automatique de la parole (RAP) basée sur des méthodes stochastiques atteint d'excellents niveaux de performances pour de nombreuses langues si des corpus d'entraînements (textuels et audio) de tailles suffisantes sont disponibles [8]. Le principal obstacle au développement de systèmes de RAP pour les langues africaines est le manque ou l'insuffisance de corpus textuels, du fait précisément de la tradition orale de ces pays et de leur récent système graphique.

Depuis l'émergence de l'Internet et du Web, et surtout grâce aux journaux électroniques, des bases de données se constituent progressivement. Des travaux ont récemment été effectués par différentes équipes de chercheurs pour la construction automatique de corpus textuels à partir d'Internet pour les langues peu dotées [5], [16]. Inspiré par ces travaux - nous avons obtenu

[1] Certains pays comme Djibouti, disposent d'archives culturelles audio enregistrées depuis 40 ans.

[2] http ://www.ethnologue.com
[3] http ://www.sil.org
[4] advanced tongue root

	Labiales	Labiodentales	Dentales	Alvéolaires	Retroflèxes	Palatales	Vélaires	Uvulaires	Pharyngales	Glottales
Occlusives voisées	b		d		dh		g	q		'
Occlusives non voisées		t				k				
Nasales	m			n						
Fricatives non voisées		f		s		sh		kh	x	h
Fricatives voisées						j			c	
Roulées				r						
Latérales				l						
Approximantes	w					y				

TAB. 1: Structure phonétique des consonnes de la langue somalienne.

pour la langue somalienne - un texte brut de 3 millions de mots issus d'articles de journaux. La table 2 montre la composition du corpus textuel. Il est composé de 2 820k mots, avec 121k mots différents.

TAB. 2: Composition du corpus textuel issu de l'Internet.

Phrases	84,7k
Mots	2 820k
Mots distincts	121k
Racines	6 042k
Racines distinctes	4,4k
Phonèmes	14 104k
Phonèmes distincts	36

3.2. *Corpus audio : Asaas*

Un sous-ensemble du corpus textuel a été isolé pour servir de base aux enregistrements sonores. Ce texte a été lu par 10 locuteurs âgés de 20 à 60 ans. Les enregistrements se sont déroulés dans un environnement non bruité. Nous avons ainsi constitué un corpus de 10 heures de parole à une fréquence d'échantillonnage de 16Khz et un codage sur 16 bits ainsi que sa transcription exacte au format Transcriber [1]. Ce premier corpus audio somalien, nommé Asaas [5], contient 59k mots et 10k mots différents. Les répartitions phonétiques de Asaas et du corpus textuel sont similaires (figure 1). Ce corpus est divisé en deux parties : 9,5 heures pour l'apprentissage et 0,5 heure pour les tests.

4. BOÎTES À OUTILS SOMALI

Afin de rendre les textes somaliens exploitables, nous avons été amenés à développer une série d'outils informatiques [10]. La langue somalienne est une langue "jeune" dans sa version écrite et présente une orthographe non standardisée. Le même mot peut se trouver écrit de différentes manières. Ces transcriptions multiples ne peuvent pas être considérées comme fausses, puisse qu'aucune standardisation ne s'est imposée à ce jour. Cependant, elles perturbent la qualité des modèles stochastiques et la robustesse des systèmes automatiques. Afin de corriger ce problème, nous avons écrit un programme qui "standardise" les transcriptions en utilisant l'orthographe la plus fréquemment rencontrée comme orthographe de référence.
La plupart des mots somaliens sont formés par la concaténation d'un nombre limité de "sous-mots", nommés "racines" dans ce papier. Leur forme est en général [3] : CVC, CVVC, CVV, VC [6], *etc.* Par exemple :
. *birlab* (un aimant) – *bir* (CVC) and *lab* (CVC) ;
. *galab* (après-midi) – *gal* (CVC) and *ab* (VC).
Cette particularité de la langue somalienne nous semble très intéressante, car les racines représentent un niveau intermédiaire entre les mots complets et les phonèmes/lettres. Nous verrons dans le chapitre "reconnaissance de la parole" qu'elles peuvent être utilisées pour la transcription. Afin d'étudier ces racines, nous avons mis au point un programme qui extrait les racines des mots somaliens. 4 400 racines ont ainsi pu être extraites.
Différents transducteurs ont également été développés pour traiter les abréviations, les dates, les nombres, etc.. qui apparaissent dans les corpus. Un phonétiseur (SOMPHON) de textes somaliens, inspiré du phonétiseur LIA_PHON [2] a également été créé.

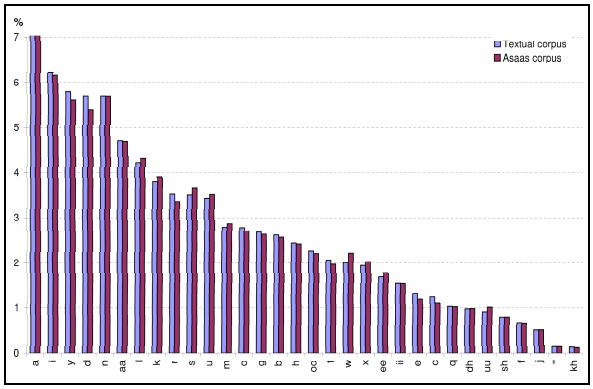


FIG. 1: Répartitions phonétiques pour le corpus textuel et le corpus audio Asaas.

5. EXPÉRIENCES

5.1. *Modélisation acoustique*

Un modèle acoustique somalien de départ utilisant un modèle français a été construit à l'aide d'une table de concordance entre les phonèmes des deux langues. Ce premier modèle a permis d'initialiser un processus itératif constitué d'une phase d'alignement puis d'apprentissage, afin d'obtenir un modèle acoustique somalien. Nous avons utilisé deux types de tables de concordance pour construire le modèle initial. La première table a été définie en utilisant les connaissances expertes des 2 sys-

[5] Asaas signifie fondation en langue somalienne

[6] C=Consonne, V=Voyelle

tèmes phonétiques. Nous avons utilisé une méthode basée sur la matrice de confusion entre les phonèmes des 2 langues pour construire la deuxième table de concordance (sans connaissance *a priori* du système phonétique de la langue cible). Les résultats obtenus avec les deux méthodes [9] sont comparables et confirment les précédents travaux [4]. Nous avons adopté une représentation en 36 modèles pour la langue somalienne. Les voyelles longues et courtes sont très différentes dans leur durée d'exécution comme le montre la figure 2. Les voyelles longues sont en moyenne 1,86 fois plus longues. Ce constat nous a amené à modéliser les voyelles longues et courtes avec des modèles acoustiques différents. L'opposition ± ATR n'a pas fait l'objet de modèles séparés.

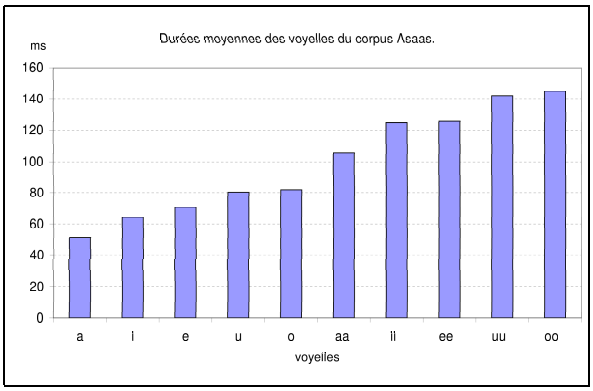


FIG. 2: Durées moyennes des voyelles du corpus Asaas. (Le rapport moyen des durées longue/courte est de 1,86)

L'analyse est effectuée sur des fenêtres de 30 ms prises toutes les 10 ms. Le signal est paramétrisé par 39 coefficients : 12 coefficients MFCC et l'énergie, plus leurs dérivées premières et secondes. Les paramètres sont centrés et réduits. Les modèles acoustiques sont composés de 3 états par phonème, excepté le " ' "[7] qui lui est codé avec 1 état, compte tenu de sa vitesse d'exécution. Nous avons utilisé pour les expériences décrits dans ce papier des modèles non contextuels avec 128 gaussiennes par état.

5.2. *Modélisation linguistique*

Un modèle de langage trigramme a été appris sur le corpus textuel somalien avec les outils du LIA et du CMU [12]. Un lexique composé des 20k mots les plus fréquents en a été extrait et a été phonétisé à l'aide de SOMPHON. Ce modèle est composé de 726k bigrammes et de 1,75M trigrammes. La perplexité, calculée sur le corpus de test, est de 63,97 avec un taux de mot hors vocabulaire (OOV) [8] de 6,77%. De la même façon, nous avons appris un modèle de langage basé sur les racines. Pour cela, le corpus de texte a été entièrement décomposé en racines. Le modèle de langage obtenu compte 4,4k racines, 189k bigrammes de racines et 996k trigrammes. L'ensemble des 4 400 racines a été retenu et phonétisé pour constituer le lexique. La perplexité calculée sur le corpus de test, lui même transformé en racines, est de 19,05 et le taux de racines hors vocabulaire est de 0,03%.

[7] Occlusive glottale.
[8] Out Of Vocabulary

5.3. *Reconnaissance de la parole*

Les premiers résultats de reconnaissance automatique de la parole en langue somalienne avec le moteur Speeral du LIA [11] ont donné un taux d'erreur mot de 20,9%. C'est un résultat encourageant, compte tenu de la petite taille de nos corpus de départ. Signalons toutefois que les mêmes locuteurs se retrouvent dans le test et l'apprentissage. La normalisation des formes orthographiques a amené un gain relatif de 34% (WER=32% avec les données non normalisées). La table 3 donne les détails des résultats obtenus.

TAB. 3: Taux d'erreur mots pour la reconnaissance de la parole en langue somalienne, avec et sans normalisation.

	Corrects	Sub	Dél	Ins	WER
Non normalisé	75,2	19,2	5,6	7,1	32,0
Normalisé	84 ;2	13,2	1,9	5,2	20,9

Le but final de notre travail n'est pas de retranscrire le plus fidèlement possible le somalien, mais de trouver un mode de représentation des données audio permettant leur indexation. C'est pourquoi, nous avons voulu évaluer les performances du système de transcription au niveau des racines. Pour cela, les hypothèses fournies par le système Speeral ainsi que les fichiers de références ont été décomposés en racines. Nous avons obtenu un taux d'erreur mot-racines [9] (WRER) de 14,2%. Ce résultat est intéressant, compte tenu de nos objectifs. Une indexation basée sur les racines et non sur les mots pourrait s'avérer plus appropriée pour les données que nous projetons de traiter. Afin d'étayer notre hypothèse, nous avons également effectué une reconnaissance basée uniquement sur les racines, en utilisant le lexique des 4 400 racines et le modèle de langage appris sur ces racines. Le taux d'erreur racines [10] (RER) est de 18,3%. La table 4 donne les détails des résultats du WRER et du RER.

TAB. 4: Taux d'erreur mots-décomposés-en-racines (WRER) et taux d'erreur racines (RER) pour la reconnaissance de la parole en langue somalienne.

	Corrects	Sub	Dél	Ins	Taux erreur
WRER	87,8	8,0	4,2	1,9	14,2
RER	83,3	10,8	5,9	1,7	18,3

Comme déjà mentionné, il est difficile pour ne pas dire impossible de trouver des corpus textuels correspondants aux périodes des données que nous souhaitons traiter. Le présent corpus, issu de l'Internet, n'est pas forcément adapté à cette tâche. Un décalage temporel et thématique est à prévoir entre les données d'apprentissage et les archives culturelles, qui se traduira, entre autres, par une augmentation des mots hors vocabulaire et une baisse du taux de reconnaissance. Les racines présentent de nombreux avantages par rapport aux mots. Etant à la base de la constitution de la langue, elles ont la capacité de la représenter

[9] Word-root error rate
[10] Root error rate

avec peu d'individus (4 400 pour la totalité de notre corpus). Ce nombre de racines augmente la représentativité d'un corpus textuel de taille limité, diminue les mots hors vocabulaires et permettra également d'accroître la portée des modèles de langage (4 ou 5 grammes). Afin de comparer la robustesse des différentes représentations, nous avons calculé les taux d'erreur mots et d'erreur racines obtenus sur deux corpus de tests différents [11]. Les résultats obtenus sont présentés dans la figure 3. Le WER augmente de 24,8%, le RER de 7,1% et le WRER de 12,6%. Le RER semble donc moins sensible par rapport au WER et également par rapport au WRER. D'autres expériences sur des corpus de période différentes devront être menées pour confirmer nos hypothèses.

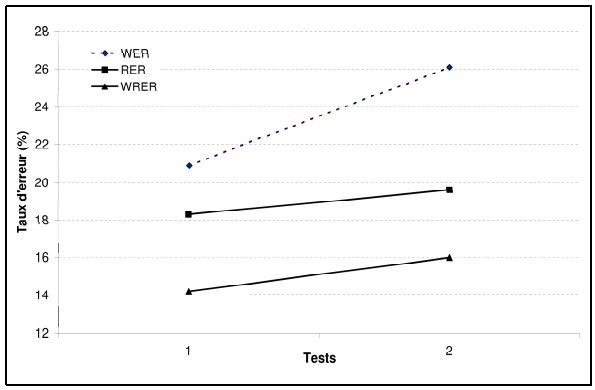


FIG. 3: Comparaison de l'évolution du WER, du WRER et du RER pour deux corpus différents.

6. CONCLUSIONS ET PERSPECTIVES

Dans ce travail, nous avons constitué un premier corpus audio de la langue somalienne. Les premiers résultats obtenus sur la reconnaissance de la parole en langue somalienne sont encourageants, compte tenu de la taille réduite de nos corpus. Nous avons aussi montré qu'un travail préalable de normalisation est nécessaire (gain relatif de 34% du WER) pour cette langue et probablement aussi pour l'ensemble des langues récemment transcrites. Nous avons également confirmé les travaux précédents concernant [16], [5] l'utilisation des documents provenant de l'Internet pour la constitution d'un corpus textuel et [4] pour la méthode rapide de modélisation acoustique. La reconnaissance automatique de la langue somalienne, basée sur les racines, semble une voie intéressante, du fait de sa moindre sensibilité aux décalages entre les données de test et d'apprentissages.

Les travaux futurs concerneront prioritairement la confirmation des résultats montrés par ces travaux par des expériences sur une plus grande échelle. Si l'utilisation des racines se confirme au niveau du traitement des données audio comme étant plus robuste, il faudra étudier l'impact d'une telle représentation plutôt que celle en mots en recherche documentaire. Enfin, nous tenterons de transposer les résultats obtenus à la langue afare, parlée à Djibouti, et qui concerne également une grande partie des données ciblées.

[11] Il s'agit uniquement d'une différence thématique dans cette expérience. Les deux corpus sont de la même période.

7. REMERCIEMENTS

Ce travail a été financé par le centre d'études et de recherches de Djibouti (CERD), le service de coopération et d'action culturelle (SCAC) du ministère des affaires étrangères français et le laboratoire informatique d'Avignon (LIA).

RÉFÉRENCES

[1] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, 1-2(33) :5–22, 2001.

[2] F. Bechet. Lia_phon : Un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 2(1) :47–67, 2001.

[3] Sabrina Bendjaballah. La palatisation en somali. *Linguistique Africaine*, (21 - 98), 1998.

[4] Huerta J.M. Khudanpur S. Marthi B. Morgan J. Peterek N. Picone J. Wang W. Beyerlein P., Byrne W. Towards language independant acoustic modeling. *IEEE workshop on automatic speech recognition and understanding*, 1999.

[5] Rayid Ghani, Rosie Jones, and Dunja Mladenic. In *Mining the web to Create Minority Language Corpora*, Berlin, 2000.

[6] Larry Hyman. Tonal accent in somali. *Studies in African linguistics*, (12) :169–203, 1981.

[7] David Le-Gac. Structure prosodique de la focalisation : cas du somali et du français, 2001.

[8] R. De Mori. *Spoken dialogues with computers*. Academic Press, 1998.

[9] A. Nimaan, P. Nocera, and J.F. Bonastre. Towards automatic transcription of somali language. In *LREC 2006*, page A paraitre, Genevo, ITALIA., 2006.

[10] A. Nimaan, P. Nocera, and J.M Torres-Moreno. Boîte à outils tal pour des langues peu informatisées : le cas du somali. In *JADT 2006 Journées d'Analyses des Données Textuelles*, page A paraitre, Besançon, FRANCE., 2006.

[11] P. Nocera, G. Linares, D. Massonie, and L. Lefort. Brno. In *Phoneme lattice based A* search algorithm for speech recognition*, TSD2002, 2002.

[12] R. Rosenfeld. The cmu statistical language modeling toolkit, and its use. In *ARPA Spoken Language Technology Workshop*, Austin, TEXAS, USA., 1995.

[13] John Saeed. *Somali reference grammar*. Dunwoody Press, MD, 1993.

[14] John Saeed. *Somali*. Johns Benjamins Publishing Company. London Oriental and African Language 10, Amsterdam/Philadelphia, 1999.

[15] Unesco. Convention pour la sauvegarde du patrimoine culturel immateriel. *http ://www.unesco.org/*, 2003.

[16] D. Vaufreydaz, M. Akbar, and J. Roullard. Asru'99. In *Internet documents : a rich source for spoken language modelling*, pages pp. 177 – 280, Keystone Colorado (USA), 1999. Workshop.

Ancres macrophonétiques pour la transcription automatique

Daniel Moraru, Guillaume Gravier

IRISA
Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
daniel.moraru@irisa.fr, guillaume.gravier@irisa.fr

ABSTRACT

Automatic speech recognition mainly rely on hidden Markov models (HMM) which make little use of phonetic knowledge. As an alternative, landmark based recognizers rely mainly on precise phonetic knowledge and exploit distinctive features. We propose a theoretical framework to combine both approaches by introducing prior (phonetic) knowledge in a non stationary HMM decoder. To demonstrate the potential of the method, we investigate how broad phonetic landmarks could be used to improve a HMM decoder by focusing the best path search. We show that, assuming error free landmark detection, every broad phonetic class brings a small improvement. The use of all the classes reduces the error rate from 22% to 14% on a broadcast news transcription task. We also experimentally validate that landmarks boundaries does not need to be detected precisely and that the algorithm is robust to non detection errors.

1. INTRODUCTION

Dans le cadre de la reconnaissance automatique de la parole par modèle de Markov caché, le décodage consiste à construire un graphe incluant l'ensemble des sources d'information disponibles (modèle de langage, lexique de prononciations, modèles acoustiques) et à effectuer une recherche du meilleur chemin dans ce graphe pour trouver la séquence de mots

$$\hat{w} = \arg\max_{w} p(y|w)p(w) \qquad (1)$$

à l'aide de l'algorithme de Viterbi. L'approche par modèle de Markov caché se base au niveau acoustique sur l'apprentissage et ne permet pas de prendre en compte des connaissances phonétiques précises. A l'inverse de cette approche, de nombreux travaux ont porté sur l'utilisation explicite de traits phonétiques pour représenter le signal de parole à des fins de reconnaissance automatique [7, 5, 4]. Les approches basées sur des traits phonétiques reposent sur une détection de traits fins comme les *onsets* et *offsets* des fricatives et des consonnes voisées [5] ou encore des traits distinctifs [7, 4]. Cependant, en pratique, la détection automatique de tels traits est délicate, notamment sur des signaux bruités ou sur de la parole spontanée.

Nous proposons une approche permettant d'introduire des connaissances (phonétiques) a priori dans le formalisme des modèles de Markov cachés et appliquons cette approche à la connaissance des macroclasses phonétiques composant le signal. Alors que les systèmes à base de traits phonétiques utilisent ces derniers comme descripteurs du signal, nous proposons de les utiliser pour guider la recherche du meilleur chemin dans le graphe de décodage. En ce sens, les connaissances a priori sont utilisées comme *points d'ancrage* de l'algorithme de décodage. Dans le cadre de cette étude, nous cherchons à valider le modèle proposé et à évaluer l'impact sur le système de reconnaissance de la parole. Dans ce but, nous utilisons des points d'ancrage déterminés manuellement. Cependant, le cadre proposé permet de prendre en compte les erreurs de détection des points d'ancrage.

Après une description du principe du décodage avec ancrage, nous étudions l'apport des différentes macroclasses phonétiques au décodage. Nous étudions ensuite l'impact de la précision de la détection des points d'ancrage macrophonétiques sur les performances avant de conclure sur des perspectives.

2. DÉCODAGE AVEC CONTRAINTES

La plupart des systèmes de transcription utilisent plusieurs passes de transcription. Chaque passe permet de générer la meilleure phrase selon l'Eq. (1) ainsi qu'un graphe de mots, ce dernier représentant de manière compact un ensemble de phrases. Chaque passe est basée sur le même principe de recherche du meilleur chemin dans un treillis en utilisant des modèles et des connaissances de plus en plus fines. En raison de la taille des treillis, une recherche exhaustive du meilleur chemin est impossible en pratique et l'on a recours à une recherche approximative dite en faisceau. Nous décrivons succinctement cet algorithme de recherche avant de l'étendre pour introduire des connaissances a priori.

2.1. Décodage en faisceau

Le décodage en faisceau se base sur l'algorithme de Viterbi pour la recherche du meilleur chemin dans un treillis. Cet algorithme procède de manière incrémentale en recherchant l'hypothèse partielle optimale menant à un instant donné dans un état (j, t) du treillis, selon

$$S(j,t) = \max_i S(i, t-1) + \ln(a_{ij}) + \ln(p(y_t|j)) \ , \qquad (2)$$

où $\ln(a_{ij})$ est le poids d'une transition de l'état i vers l'état j et $p(y_t|j)$ la vraisemblance d'un descripteur y_t conditionnellement à l'état j. $S(i,t)$ représente le score du meilleur chemin partiel menant à l'état i à l'instant t.

Dans la recherche en faisceau, à chaque instant, seules les hypothèses les plus prometteuses, vérifiant

$$S(i,t) \geq \max_j S(j,t) - \alpha \ , \qquad (3)$$

sont conservées, α étant un facteur d'élagage contrôlant

la taille du faisceau. En général, on fixe également un nombre maximum d'hypothèses à chaque instant.

2.2. Introduction de points d'ancrage

Nous étendons le formalisme de l'algorithme de Viterbi pour y introduire des contraintes liées à une connaissance a priori sur le meilleur chemin. Par exemple, si l'on sait a priori qu'une portion donnée d'un segment correspond à une voyelle, il est possible de pénaliser, voire de supprimer, les chemins dans le treillis qui ne concordent pas avec cette connaissance.

D'un point de vue formel, une approche possible consiste à considérer le graphe de décodage comme non stationnaire. L'idée est que, si une transition amène dans un état du treillis (i, t) incompatible avec les connaissances a priori, alors le coût de cette transition augmente. Pour refléter la présence d'une connaissance a priori, vue comme une contrainte sur les chemins dans le procédure de décodage, nous remplaçons dans (2), les poids de transition a_{ij} par

$$\ln(a_{ij}(t)) = \ln(a_{ij}) - \lambda(t) I_j(t) \quad (4)$$

où $I_j(t)$ est une indicatrice valant 0 si le noeud (j, t) du treillis est en accord avec la connaissance a priori dont on dispose. La pénalité $\lambda(t) \geq 0$ infligée à une transition incompatible avec nos connaissances a priori dépend de la confiance accordée à ces connaissances. Par exemple, si $\lambda(t) = \infty$, on accorde alors une confiance totale aux connaissances a priori disponibles et la recherche du meilleur chemin se limitent aux seuls chemins valides par rapport aux a priori. Si $\lambda(t) = 0$, on retrouve alors un décodage sans a priori. Dans la suite de ce travail, nous considérons des a priori fiables car extraits manuellement et posons donc $\lambda(t) = \infty, \forall t$.

Dans cette approche, les connaissance a priori sont utilisées comme comme *points d'ancrage* ou *ancres* pour guider la recherche vers les solutions plausibles par rapport aux connaissances disponibles. Dans le cas extrême où $\lambda = \infty$, on peut également parler de *contraintes* sur l'espace des solutions. Soulignons que, à la différence des décodeurs à base de traits phonétiques, la non détection de points d'ancrage n'a que peu d'impact sur le décodage.

3. ANCRES MACROPHONÉTIQUES

Dans un premier temps, nous exploitons le formalisme défini précédemment pour étudier l'apport d'une connaissance sur le contenu macrophonétique du signal à reconnaître. Nous considérons les cinq macroclasses phonétiques définis par SAMPA : voyelles, occlusives, fricatives, nasales et *glides*, cette dernière classe regroupant semi-voyelles, latérales et vibrantes. L'idée est de segmenter le signal selon ces cinq classes et d'utiliser le résultat de cette segmentation pour obtenir des points d'ancrage afin de ne garder que les hypothèses valides par rapport à ces points d'ancrage. Dans la mesure où il est illusoire de détecter précisément les frontières dans la segmentation, nous n'utilisons que la partie centrale d'un segment comme ancre. Cette étude ayant pour but principal de déterminer si l'information macrophonétique est pertinente, nous utilisons des points d'ancrage déterminés manuellement.

Plusieurs considérations justifient le choix d'une information macrophonétique. D'une part, nous avons remarqué que, malgré l'élagage, la liste des hypothèses concurrentes à un instant donnée contient des hypothèses correspondant à différentes classes phonétiques. D'autre part, si la détection automatique de traits phonétiques fins est délicate, la détection/segmentation de macroclasses phonétiques donnent de bonnes performances [1, 3, 6].

3.1. Corpus

Les expériences sont réalisées sur un corpus de quatre heures d'émissions radiophoniques, extrait du corpus de développement de la campagne ESTER [2]. Le corpus est essentiellement composé de parole planifiée de haute qualité avec quelques entrevues comportant de la parole plus spontanée et/ou des conditions acoustiques dégradées. Afin de dériver des points d'ancrage, nous avons réalisé un alignement phonétique forcé du corpus. Pour une macroclasse donnée, les points d'ancrage sont obtenus à partir de l'alignement phonétique de référence. Pour chaque phone appartenant à la classe considérée, on génère une ancre centrée sur le phone et dont la durée est proportionnelle à celle du phone. Dans cette première expérience, la durée d'une ancre est fixée à 50% de la durée du phone correspondant.

3.2. Systèmes de transcription

Nous utilisons les points d'ancrage dans deux systèmes de transcriptions. Pour les deux systèmes, une première passe permet de générer un graphe de mots à l'aide d'un modèle de langage trigrame, le graphe de mots étant ensuite réévalué avec des modèles de phones contextuels inter-mots. La différence entre les deux systèmes résident dans la finesse des modèles acoustiques utilisés pour la première passe : des modèles hors contexte dans un cas et des modèles contextuels intra-mots dans l'autre. Les modèles hors-contexte possèdent 114 états à 128 gaussiennes tandis que les modèles contextuels possèdent 4 019 états à 32 gaussiennes.

La taille du treillis de la deuxième passe est relativement petite puisque limitée par le graphe de mots. En revanche, le treillis exploré lors de la première passe est constitué par la composition des graphes représentant le modèle de langage, les prononciations et les modèles acoustiques. Nous considérons donc l'introduction de points d'ancrage au niveau de la première passe afin de faciliter la recherche du meilleur chemin dans le treillis le plus complexe. L'utilisation de points d'ancrage lors de la première passe vise également à limiter la taille du graphe de mots généré.

3.3. Résultats

Le tableau 1 regroupe les taux d'erreurs obtenus pour les deux systèmes avec chacune des classes phonétiques étudiées. Nous reportons également la proportion de signal correspondant à un point d'ancrage (signal/ancre).

L'utilisation de l'ensemble des points d'ancrage permet une amélioration très nette du taux d'erreur après chacune des passes, aussi bien pour les monophones que pour les triphones. Le gain de performance est évidemment plus conséquent lorsque l'on utilise des monophones, ces derniers donnant lieu à plus de confusions phonétiques lors d'un décodage acoustico-phonétique. L'introduction d'une connaissance supplémentaire permet alors de lever plus d'ambiguïtés que pour les triphones. En revanche,

TAB. 1: Taux d'erreur de transcription (en %) après chaque passe pour les différentes classes d'ancrage.

ancres		aucune	toutes	voy.	occ.	fri.	nas.	gli.
signal/ancre		0,0	43,6	18,3	9,0	7,3	2,8	6,2
monophones	passe 1	29,2	15,3	26,6	26,5	27,5	27,8	25,1
	passe 2	22,3	13,9	21,2	20,7	21,0	21,5	20,1
triphones	passe 1	27,3	19,6	27,0	26,3	26,0	26,4	24,9
	passe 2	21,3	15,0	20,7	20,4	20,3	20,7	19,6

il est intéressant de constater que les deux systèmes obtiennent des taux d'erreur comparables, voire légèrement meilleur pour le système monophone, après la deuxième passe. Soulignons également le faible taux d'erreur obtenu avec les monophones après la première passe, comparable au taux d'erreur avec les triphones après les deux passes. Ce résultat suggère que l'utilisation de points d'ancrage en conjonction avec des modèles phonétiques simples permet une nette amélioration des performances pour des temps de décodage faible, le nombre moyen d'hypothèses actives par trame étant divisé par 4 avec les points d'ancrage.

Si l'on regarde l'influence des points d'ancrage de chacune des classes phonétiques, il apparaît clairement que chaque classe apporte une information supplémentaire. En effet, les classes considérées séparément apportent chacune un léger gain de performances, le plus grand gain étant obtenu avec les *glides*. Cette dernière classe correspond à des phonèmes transitoires, fortement affectés par la coarticulation, pour lesquels les modèles acoustiques font de nombreuses confusions. Le gain obtenu avec des points d'ancrage sur les voyelles est plus surprenant dans la mesure où les modèles des deux systèmes de reconnaissance font peu de confusions entre les voyelles et les autres classes. Ceci s'explique par le fait que la liste des nœuds du treillis actifs à un moment donné dépend non seulement du score $p(y_t|i)$ de la trame courante mais également de tout le chemin parcouru jusqu'à ce nœud. Il se peut donc que même si, localement, les probabilités $p(y_t|i)$ sont plus élevés pour les nœuds i correspondant à une voyelle, certains chemins incompatible avec le point d'ancrage ait un bon score global. Il est probable qu'une technique d'anticipation phonétique (phone look-ahead) permettrait de limiter cet effet.

L'utilisation de points d'ancrage lors de la première passe du système de transcription permet, outre une amélioration du taux d'erreur, de limiter la taille du graphe de mots produit. Le tableau 2 indique la taille moyenne des graphes de mots obtenus pour chaque classe de points d'ancrage avec les monophones et les triphones. La taille des graphes est donnée en nombre d'arcs et de nœuds par trames, avec un débit de 100 trames par seconde. Les nœuds du graphe correspondent aux frontières de mots tandis que les arcs correspondent aux mots. Nous donnons également le taux d'erreur oracle (GER) après alignement du graphe avec la transcription de référence. Les conclusions sur la taille des graphes sont similaires en tout point à celles énoncées précédemment sur les taux d'erreur. L'ensemble des points d'ancrage permettent une forte réduction de la taille des graphes de mot, réduction plus marquée pour les monophones qui produisent des graphes de mots plus grands en l'absence de contraintes. Chaque classe d'ancrage apporte une réduction du graphe de mot, les *glides* offrant la plus grande réduction.

Ces résultats montrent que l'utilisation de l'ensemble des ancres macrophonétiques permettrait d'une part de faire grandement diminuer le taux d'erreur des systèmes de transcription, alors que l'apport de chacune de ces classes d'ancrage est minime, et, d'autre part, d'accélérer le décodage. Dans la mesure où les points d'ancrage limitent le faisceau de décodage, il s'avère avantageux de les utiliser avec des modèles acoustiques peu sélectifs qui présentent a priori moins de conflits avec les points d'ancrage.

4. PRÉCISION DES POINTS D'ANCRAGE

Dans les expériences précédentes, nous utilisons des points d'ancrage représentant 50% de la durée des phones correspondant, c'est-à-dire qu'un [a] d'une durée de τ secondes donnera lieu à une ancre de type voyelle d'une durée de $\tau/2$ secondes, centrée sur le milieu du phone. Nous étudions dans cette partie l'importance de la précision temporelle des points d'ancrage en variant la durée des points d'ancrage de 10% à 70% de la durée du phone. Le tableau 3 regroupe les résultats obtenus pour les triphones avec l'ensemble des points d'ancrages, en terme de taux d'erreur après chacune des passes (WER 1 et WER 2), de taille de l'espace de recherche (# hyps, le nombre moyen d'hypothèses actives par trame) et de qualité du graphe de mot généré par la première passe. Ces résultats montrent que, bien évidemment, lorsque la taille des points d'ancrages augmente, le taux d'erreur diminue ainsi que le nombre d'hypothèses propagées et la taille du graphe de mot. De manière plus surprenante, on note que des points d'ancrages d'une durée égale à 10% de la durée des phones permettent déjà une très nette amélioration des performances ainsi qu'une réduction sensible du nombre moyen d'hypothèses actives et de la taille des graphes de mots. Notons que dans cette configuration, les points d'ancrage correspondent seulement à 8,7% de la durée totale du signal. La même expérience en n'utilisant que les voyelles comme points d'ancrage montre une tendance similaire avec une baisse du taux d'erreur de 20,7% à 19,5% pour une durée des points d'ancrage de 10% (soit 3,6% de la durée totale du signal correspondant à des points d'ancrage).

Ces résultats établissent clairement que la détection des points d'ancrage n'a pas besoin d'être précise puisque, avec seulement 10% de la durée réelle, le taux d'erreur baisse de manière significative. Une plus grande précision temporelle dans la détection des ancres permet une légère amélioration des performances et surtout une accélération du décodage en limitant le nombre de chemins explorés dans le treillis de décodage.

Nous avons également pu montré la robustesse de l'approche proposée face aux erreurs de non détection des ancres, en simulant ce type d'erreurs pour différents taux de faux rejet. Sur une heure d'émission, le taux d'erreur

TAB. 2: Taille et taux d'erreur des graphes de mots générés avec les différentes classes d'ancrage.

	monophones			triphones		
contraintes	# arcs	# nœuds	GER (%)	# arcs	# nœuds	GER (%)
aucune	54,8	8,2	7,9	37,0	6,9	7,5
toutes	23,4	4,0	3,0	24,3	4,8	5,0
voyelles	43,6	7,0	7,3	36,4	6,8	7,7
occlusives	45,9	7,1	7,0	36,5	6,8	7,0
fricatives	47,5	7,3	7,2	35,5	6,7	6,9
nasales	49,7	7,6	7,3	35,7	6,7	7,2
glides	39,1	6,3	6,6	31,8	6,1	6,9

TAB. 3: Taux d'erreur, tailles des graphes de mots et de l'espace de recherche en fonction de l'étendue des points d'ancrage. Les résultats sont données sur une heure d'émission (France-Info), pour le système triphones avec l'ensemble des points d'ancrage.

étendue (en %)	0	10	20	30	40	50	60	70
WER 1 (%)	26,3	19,9	19,4	18,8	18,8	18,5	18,8	20,0
WER 2 (%)	20,7	15,2	15,1	15,0	14,8	14,4	14,3	14,7
# hyps	51 730	31 346	29 444	27 240	24 813	22 421	20 091	18 311
# arcs	38,1	30,5	29,8	28,7	26,8	24,9	22,7	20.96
# nœuds	7,0	5,9	5,8	6,0	5,2	4,9	4,5	4,2
GER (%)	6,5	4,5	4,5	4,5	4,6	4,5	4,7	5,1

pour le système monophone est de 13.9% avec l'ensemble des ancres et de 22.3% sans points d'ancrage. Ce taux d'erreur est respectivement de 15,8% et 17,9% pour des taux de non détection des ancres de 25% et de 50%.

5. DISCUSSION

Ce travail propose un cadre théorique pour l'introduction de connaissances a priori dans un système de décodage basé sur la recherche d'un chemin optimal dans un treillis. Nous appliquons ce cadre à la reconnaissance de la parole en introduisant dans le décodage une connaissance oracle des macroclasses phonétiques présentes dans le signal. Les résultats montrent que cette simple information permet une très nette amélioration des performances en transcription, même en conjonction avec des modèles de phones peu performants. De plus, nous mettons en évidence qu'il n'est pas nécessaire de détecter les points d'ancrage avec une bonne précision temporelle. Ces premiers résultats montrent donc l'intérêt de travailler sur un détecteur fiable d'ancres macrophonétiques.

La première perspective de ce travail est donc l'étude du comportement de cette approche avec une détection automatique des points d'ancrage, notamment en ce qui concerne la robustesse aux erreurs de la détection automatique. Une piste de travail consiste à exploiter des mesures de confiances sur la détection des points d'ancrage pour adapter la pénalité $\lambda(t)$ en conséquence. Nous pensons également que l'utilisation d'ancres macrophonétiques devrait permettre une robustesse accrue par rapport aux changement de condition acoustique ou du style de parole, dans la mesure où l'on est en droit d'attendre plus de robustesse d'une segmentation macrophonétique que de modèles de phones.

Une deuxième perspective réside dans l'étude d'autres points d'ancrage, comme les traits distinctifs. Enfin, dans le strict cadre de la reconnaissance de la parole, l'intégration de connaissance phonétique est possible à d'autres niveaux, par exemple à travers un choix dynamique du vocabulaire et/ou du modèle de langage ou encore un élagage du graphe de mots. Notre approche réalise d'une certaine manière une stratégie d'adaptation dynamique du modèle de langage en pénalisant les mots incompatibles avec les points d'ancrage.

Pour conclure cette discussion, notons que le cadre générique proposé s'applique à n'importe quel problème pouvant se formuler sous la forme d'un décodage par optimisation dans un graphe et possèdent donc des applications dans de nombreux domaines autre que la reconnaissance de la parole.

RÉFÉRENCES

[1] M. Chen. Nasal landmark detection. In *Intl. Conf. Speech and Language Processing*, pages 636–639, 2000.

[2] S. Galliano, E. Geoffrois, J.-F. Bonastre, G. Gravier, D. Mostefa, and K. Choukri. Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news. In *Language Resources and Evaluation Conference*, 2006. to appear.

[3] A. Howitt. Vowel landmark detection. In *Intl. Conf. Speech and Language Processing*, 2000.

[4] John Hopkins University, Center for Language and Speech Processing. *Landmark-based speech recognition : report of the 2004 John Hopkins Summer Workshop*, 2005.

[5] A. Juneja. *Speech recognition based on phonetic features and acoustic landmarks*. PhD thesis, University of Maryland, 2004.

[6] J. Li and C.-H. Lee. On designing and evaluating speech event detectors. In *European Conf. on Speech Communication and Technology*, 2006.

[7] S. A. Liu. *Landmark detection for distinctive feature-based speech recognition*. PhD thesis, Massachusetts Institute of Technology, 1995.

Session IV

Poster

Lundi 12 juin 2006 - 16h45 18h00

Détection et correction automatique des déviations dans la réalisation de l'accent lexical anglais par des apprenants français

Guillaume HENRY, Anne BONNEAU, Vincent COLOTTE

Speech Group
LORIA/CNRS and INRIA
Campus Scientifique - BP 239
Vandoeuvre-lès-Nancy 54506, France
{Guillaume.Henry, Anne.Bonneau, Vincent.Colotte}@loria.fr

ABSTRACT

The work presented here is developed within a project devoted to the acquisition of English prosody by French learners, using speech technology modifications, and knowledge about L1 and L2 prosody. Our goal is to provide learners with relevant feedback in an automatic manner by comparing the prosodic cues of their realizations to that of a model. We focus on the English lexical accent and propose methods to correct automatically the learner's realizations. We present our strategy and illustrate it through a concrete example.

1. INTRODUCTION

Nos travaux se placent dans le cadre de l'action « Assistance à l'apprentissage des langues » du Plan Etat Région et concernent plus particulièrement l'apprentissage de la prosodie anglaise par des apprenants français.

Depuis quelques années, nous voyons l'émergence de logiciels dédiés à l'apprentissage de la composante orale des langues assisté par ordinateur. Parmi ceux-ci, Winpitch LTL [10], utilisé par des enseignants de langue de l'université, propose une visualisation en temps réel de la courbe mélodique de l'apprenant et possède des fonctions de modifications du signal. BetterAccent [8] propose une comparaison visuelle des tracés prosodiques des apprenants avec ceux d'une référence et indique, à partir de la réalisation modèle, les indices que l'apprenant doit reproduire correctement. Le module prosodique de SLIM [5] propose une évaluation automatique de la durée relative des syllabes au sein d'un mot, afin d'améliorer la réalisation (par un locuteur italien) de l'accent lexical anglais (comparaison avec une référence humaine). Cependant, si on excepte le module d'analyse de la durée de SLIM, ces logiciels ne proposent pas d'évaluation automatique de la réalisation de l'apprenant.

Notre but est de fournir aux apprenants non seulement une visualisation des indices prosodiques, mais également une évaluation automatique de sa production, ainsi qu'une correction auditive. Pour ce faire, nous exploitons des outils d'analyse et de traitement du signal développés dans notre équipe ainsi que des connaissances sur la prosodie de la langue maternelle (L1) et de la langue seconde (L2). L'évaluation est donnée sous la forme d'indices visuels et de petits textes.

Nous avons choisi de nous concentrer dans un premier temps sur la réalisation de l'accent lexical. Ce papier présente notre démarche, que nous résumons ci-dessous, ainsi que quelques tests. Pour établir une évaluation, la production de l'apprenant est comparée à un modèle (provenant d'une référence humaine, pour l'instant). Après une segmentation en syllabes et en phonèmes des deux productions, l'accent lexical est automatiquement localisé. Ensuite, une comparaison entre les indices prosodiques de l'apprenant français et les indices prosodiques du modèle est réalisée dans le but d'affiner l'évaluation. Enfin, on procède à la correction automatique des réalisations. Il s'agit en fait de modifier automatiquement les paramètres prosodiques de la réalisation de l'apprenant dans le but de s'approcher le plus possible de la réalisation du modèle, tout en gardant le timbre et la hauteur de voix de l'apprenant.

Nous présentons tout d'abord le corpus et les outils dont nous disposons (partie 2), puis la méthode utilisée (partie 3). Enfin, nous proposons une première évaluation de cette méthode (partie 4).

2. CORPUS ET OUTILS

2.1. Corpus

Dans le cadre du Plan Etat-Région, nous sommes en collaboration avec des enseignants d'anglais de l'université de Nancy 2 et du secondaire. Ce partenariat a débouché sur la mise au point d'une base d'exercices progressifs, et d'un corpus associé. Le corpus est composé de mots isolés transparents, de phrases (quelques centaines) et de petits textes (quelques dizaines). Il a été enregistré par deux enseignants d'anglais (un homme et une femme) de langue maternelle anglaise. Nous avons sélectionné dans un premier temps une liste de mots isolés transparents qui mettent bien en lumière les problèmes d'accentuation des locuteurs français.

2.2. Outils de traitement du signal dédiés à l'apprentissage des langues

Des fonctions de modification du signal de parole utilisant une version améliorée de TD-PSOLA [4] ont été

développées et incluses dans Winsnoori, logiciel dédié à la visualisation, au traitement et à l'analyse de la parole [9]. Concrètement, l'utilisateur a la possibilité de modifier le contour mélodique d'une phrase ou d'une partie d'une phrase, de resynthétiser le signal et de sauvegarder la modification effectuée. Il en est de même avec le débit : l'utilisateur peut non seulement appliquer un ralentissement du débit de la parole afin d'améliorer l'intelligibilité du signal, mais aussi modifier la durée de chaque segment. Ces deux traitements peuvent s'effectuer soit simultanément, soit séparément. Ils permettent aux utilisateurs de modifier les paramètres prosodiques de leur réalisation afin de se rapprocher du modèle tout en gardant leur timbre et leur registre de voix. Cette correction semble bénéfique dans le processus d'apprentissage d'une langue étrangère [2].

3. METHODE

3.1. Exploitation de la prosodie de L1 et L2

Lors du processus d'apprentissage d'une langue étrangère, la langue maternelle (L1) influence considérablement les réalisations des apprenants dans une langue seconde (L2) [1]. Cette influence est particulièrement importante pour les langues anglaises et françaises qui sont classées dans deux catégories prosodiques différentes, le français étant considérée comme « syllable-timed » et l'anglais comme « stress-timed ».

L'évaluation que nous proposons exploite donc les difficultés spécifiques des apprenants français dans leur réalisation de la prosodie de l'anglais. La réalisation de l'accent lexical anglais est particulièrement difficile pour les français puisque sa place est libre, alors que celle du français est fixe, et qu'il est très marqué acoustiquement, contrairement à l'accent français. Plus précisément, l'accent lexical français est essentiellement caractérisé par un allongement de la durée de la dernière syllabe du mot ou d'un groupe de mots. L'accent lexical anglais est généralement marqué par une forte augmentation de l'intensité, accompagnée d'une modification de la hauteur et d'un allongement de la durée des noyaux vocaliques.

En outre, en anglais, les voyelles des syllabes inaccentuées sont souvent réduites. Ceci est une caractéristique de la langue anglaise que les locuteurs français ont du mal à assimiler puisqu'en français les voyelles inaccentuées conservent leur timbre. Enfin, les occlusives sourdes sont aspirées sous l'accent en anglais.

Lors de la réalisation de l'accent lexical anglais, un français va avoir tendance à garder ses propres schémas prosodiques, et à atténuer les caractéristiques de l'accent anglais. On peut ainsi prédire les principales déviations de l'apprenant. Le locuteur français aura tendance à allonger la dernière syllabe du mot, même lorsque celle-ci ne doit pas recevoir d'accent lexical. En admettant qu'il ait accentué la bonne syllabe, il est probable que les indices prosodiques ne seront pas suffisamment marqués : la syllabe à accentuer ne sera ni suffisamment intense ni suffisamment longue par rapport aux autres, les différences de hauteur seront trop faibles. La prononciation française sera également caractérisée par une absence de réduction et d'aspiration (occlusives sourdes sous l'accent).

En plus des problèmes d'accentuation, les apprenants français ont tendance à mal syllabifier les mots anglais, ce qui rend les comparaisons entre les réalisations anglaises et françaises plus complexes.

3.2. Evaluation automatique

L'effet bénéfique des retours visuels dans l'apprentissage des langues n'est plus à prouver [3]. Dans notre logiciel, les courbes mélodiques et énergétiques, ainsi que les durées de chaque segment et chaque syllabe sont représentées sur le spectrogramme.

La stylisation du contour mélodique est linéaire [7] et l'écart de hauteur entre la syllabe accentuée et les autres syllabes du mot est évalué en demi-tons.

Pour analyser correctement les retours visuels, il faut au préalable segmenter les différentes réalisations en syllabes et en phonèmes. L'étudiant prononce un mot ou un texte tiré du corpus (dans un premier temps, nous nous sommes limités au mot). La segmentation en phonèmes des réalisations des locuteurs anglais (natifs) est réalisée grâce à un alignement texte-parole développé au sein de l'équipe et utilisant des modèles de Markov [6]. En ce qui concerne les locuteurs non-natifs, l'alignement texte-parole nécessite un apprentissage spécifique afin d'adapter les modèles. Cet alignement est en cours de réalisation dans notre équipe. Pour la syllabation des mots isolés, nous utilisons un dictionnaire en ligne qui fournit le découpage en syllabes. Pour les phrases, des règles et des algorithmes de syllabation seront utilisés.

Une fois le signal segmenté en syllabes et en phonèmes, nous proposons une localisation de l'accent lexical. Signalons qu'il n'y a pas toujours véritablement de réalisation d'un accent lexical anglais chez l'apprenant français. Notre méthode vise alors à aider l'apprenant à se rapprocher au mieux des indices prosodiques du modèle.

Afin de détecter la position de l'accent lexical sur des mots isolés, nous avons choisi dans un premier temps de considérer comme syllabe accentuée la syllabe qui porte le pic de F0. Cet indice est efficace pour des mots isolés prononcés sans intonation particulière. Nous donnons les résultats de quelques tests sur la localisation de l'accent en section 4. Il faudra bien entendu utiliser par la suite des critères plus complexes, qui prennent en compte en particulier l'intensité.

Une évaluation plus complète est en cours de réalisation. Elle inclut des appréciations sur la réalisation des différents paramètres prosodiques, et exploite les connaissances sur la prosodie de L1 et L2. En particulier, on recherche si les différences de hauteur entre les syllabes sont suffisamment marquées, ou si la dernière

syllabe du mot n'est pas trop longue alors qu'elle n'est pas sous l'accent (erreur typique d'un locuteur français).

Nous proposons ensuite une comparaison visuelle des indices prosodiques (figure 1, deuxième spectrogramme). On indique sur la réalisation de l'apprenant la position de son accent lexical (rectangle vert) et la position de l'accent lexical de la référence (rectangle rouge). Si les deux rectangles sont superposés, le locuteur a placé l'accent sur la bonne syllabe. On montre également à l'apprenant les durées des syllabes et des voyelles du modèle (en haut du spectrogramme) et on indique si la syllabe accentuée se distingue suffisamment des autres syllabes (flèche et texte inséré). Cette évaluation est complétée par un fichier texte dans lequel sont interprétées les informations visuelles fournies.

Le phénomène de réduction, quand il va jusqu'à la suppression d'une syllabe, nous pose un problème spécifique. En effet, dans ce cas, le nombre de syllabes des réalisations des locuteurs anglais et français diffèrent (le locuteur français a tendance à prononcer toutes les syllabes). Ceci n'est pas considéré comme une erreur car la suppression de syllabes n'est pas obligatoire en anglais. Mais ce phénomène rend la comparaison difficile, voire impossible. C'est pourquoi la solution envisagée actuellement consiste à faire recommencer l'apprenant (après lui avoir indiqué qu'il avait prononcé une syllabe supplémentaire par rapport au modèle anglais) afin que la comparaison soit réalisable.

3.3. Correction automatique des déviations

Dans une version antérieure de notre logiciel, les modifications du signal de parole étaient manuelles, donc réalisées par l'utilisateur. Nous avons mis au point un retour auditif automatique qui remplace les indices prosodiques de l'apprenant par ceux de la référence tout en gardant son registre et son timbre de voix.

Grâce aux techniques de modifications du débit et de la fréquence fondamentale (F0) de la parole développées au sein de l'équipe, nous avons la possibilité de « copier » automatiquement les caractéristiques prosodiques d'un modèle. Dans un premier temps, on aligne les durées relatives des phonèmes de l'apprenant sur les durées relatives des phonèmes du modèle. Dans un second temps, on recalcule le contour mélodique de l'apprenant par une interpolation linéaire du contour du modèle. L'alignement des durées permet d'éviter les effets de bord et en particulier les problèmes de voisement/non-voisement. Ces opérations doivent sauvegarder le timbre et la hauteur de voix de l'apprenant. Pour ce faire, on procède à un recalage du contour mélodique de l'apprenant en jouant sur les moyennes de F0 de l'apprenant et de la référence. Un exemple est donné sur la figure 1 (troisième spectrogramme).

Nous envisageons également une autre stratégie de correction : le renforcement automatique des déviations les plus importantes commises par l'apprenant. En exagérant ces déviations, on espère que l'apprenant prendra conscience de ce qu'il ne faut pas faire.

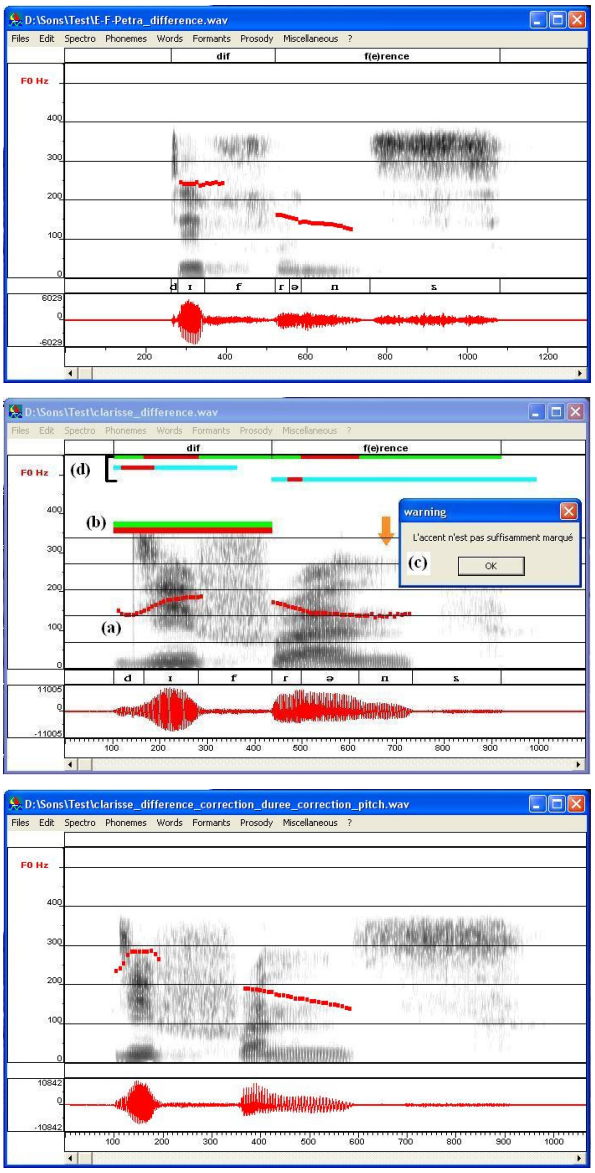


Figure 1 : *Comparaison de la prononciation du mot « difference » : en haut la référence, au centre l'apprenant français et en bas la correction automatique proposées à partir de la voix de l'apprenant. Au centre, la courbe (a) représente le contour mélodique original superposé sur le spectrogramme ; les barres horizontales (b) indiquent les positions des accents ; un petit texte (c) affine l'évaluation, et les durées relatives (d) sont mises à l'échelle de l'apprenant (en haut et en vert pour l'apprenant, en bas et en bleu pour la référence).*

3.4. Un exemple

Nous analysons ici le mot « difference » (« différence » en français) et dans le cas étudié, la référence (une locutrice anglaise) n'a pas prononcé la deuxième voyelle (« ə »).

L'interface Winsnoori nous offre la possibilité d'étiqueter le mot en syllabes et en phonèmes. La transcription

phonétique en API est la suivante : ['d ɪ f r ə n s]. La comparaison effectuée est montrée sur la figure 1.

Dans cet exemple, l'accent est bien positionné mais on voit nettement que la différence de hauteur entre les deux syllabes est forte chez la locutrice anglaise, alors qu'elle est très faible chez l'apprenant (une locutrice française). On remarque également d'importants écarts dans les durées relatives des voyelles chez les deux locutrices. Ceci montre bien que l'apprenant français a conservé les caractéristiques prosodiques de sa langue maternelle (allongement de la dernière syllabe du mot).

4. PREMIÈRE ÉVALUATION DE LA MÉTHODE

4.1. Identification de l'accent lexical

Afin de valider notre méthode (« grossière » pour l'instant) de localisation de l'accent lexical, nous avons effectué une expérience préliminaire avec les mots isolés transparents extraits de notre corpus et prononcés par une locutrice anglaise. Nous comparons en fait la position de l'accent donnée par un dictionnaire de référence (Le ROBERT & COLLINS) avec celle donnée par notre méthode. La détection est correcte pour 40 mots sur 44. Pour deux des quatre mots restants, la locutrice a accentué le mot de façon différente de celle proposée par notre dictionnaire, mais l'accent qu'elle a réalisé a été indubitablement bien localisé par notre méthode. Signalons du reste que l'accentuation choisie par la locutrice est proposée par d'autres dictionnaires. Une détection erronée provient d'un problème de frontière syllabique, et une autre d'un conflit entre indices prosodiques (l'intensité aurait été pour ce cas un meilleur indice que la F0). Ce type de problèmes sera probablement éliminé par la suite avec des critères de localisation plus complexes.

4.2. Réalisation d'une locutrice française

Les mots étudiés ci-dessus ont été prononcés par une locutrice française. Leur analyse confirme les problèmes listés précédemment (section 3.1). Les phénomènes de réduction extrême (disparition d'une syllabe), observés chez la locutrice anglaise, mais jamais constatés pour la locutrice française, ainsi que des problèmes de syllabation incorrecte, se sont avérés très fréquents.

5. CONCLUSION ET PERSPECTIVES

Dans ces travaux, nous avons proposé une évaluation automatique des déviations dans le cadre de l'apprentissage de la prosodie anglaise par des apprenants français qui exploite à la fois des outils d'analyse et de traitements du signal, et des connaissances sur la prosodie de L1 et L2.

Un travail important doit être mené pour affiner la localisation de l'accent (notamment par une prise en compte de l'énergie, et une meilleure définition des fenêtres de recherche du pic de F0). Nous devons également compléter notre évaluation, en prenant en compte plusieurs critères d'appréciation (décrits en section 3.2). Une étude sur les seuils de déclenchement des différents retours doit être conduite afin de déterminer en particulier à partir de quand une réalisation s'éloigne de la cible (par exemple quand peut on dire que la différence de hauteur entre la syllabe accentuée et les autres syllabes est significative ?). Enfin, nous projetons d'analyser des phrases simples pour lesquelles nous testerons des modèles prosodiques.

BIBLIOGRAPHIE

[1] P. C. Bagshaw. Automatic prosodic analysis for computer-aided pronunciation teaching. *PhD thesis.* University of Edinburgh, 1994.

[2] A. Bonneau, M. Camus, Y. Laprie, and V. Colotte. A computer-assisted learning of English prosody for French students. In *Proceedings of InSTIL/ICALL.* Venice 17-19 June, 2004.

[3] K. De Bot. Visual feedback on intonation I: Effectiveness and induced practice behaviour. In *Language and Speech*, volume 26, 4, 331-350, 1983.

[4] V. Colotte and Y. Laprie. Higher pitch marking precision for TD-PSOLA. In *Proceedings of XI European Signal Processing Conference (EUSIPCO)*. Toulouse, 2002.

[5] R. Delmonte. A prosodic module for self-learning activities. In *Speech Prosody.* Aix-en-Provence, 2002.

[6] D. Fohr, J.F. Mari, J.P. Haton. Utilisation de modeles de markov pour l'etiquetage automatique et la reconnaissance de BREF80. In *Journées d'études sur la parole (JEP)*, Avignon, 1996.

[7] J. 't Hart. F0 stylization in speech: straight lines versus parabolas. In *Journal of the Acoustical Society of America*, 6, 3368-3370, 1991.

[8] J. Komissarchik and E. Komissarchik BetterAccent Tutor – Analysis and Visualization of Speech Prosody. In *Proceedings of InSTIL*. Dundee, Scotland, August 2000.

[9] Y. Laprie. Snoori, a software for speech sciences. *MATISSE*, 1999.

[10] P. Martin. WinPitch LTL II, a Multimodal Pronunciation Software. In *Proceedings of InSTIL/ICALL.* Venice 17-19 June, 2004.

Perception de la colère dans un corpus de français spontané par des apprenants portugais et tchèques

Sophie de Abreu, Catherine Mathon, Daniela Perekopska

EA333 "Atelier de Recherches sur la Parole"
Université Paris 7 – Denis Diderot
2 place Jussieu, 75251 Paris Cedex 05 France
{deabreu ; mathon ; perekopska}@linguist.jussieu.fr

ABSTRACT

The aim of this paper is to show how prosody can provide a sufficient amount of information which allows recognizing the emotion of anger in a French spontaneous corpus for foreign learners of French. We present the results of a perception test carried on two groups of foreign learners of French, Portuguese and Czech. They had to listen to French sentences and evaluate the presence, or lack thereof, and the degree of anger in these sentences. We chose to use a real spontaneous corpus in order to keep the intonation of emotions in French intact. The semantic content was neutralised to put aside the information given by the content.

1. INTRODUCTION

L'étude que nous présentons ici est à l'intersection de deux domaines de recherche : l'émotion et ses manifestations vocales d'une part et l'enseignement apprentissage du français langue étrangère d'autre part. La communication relie ces deux domaines. C'est une des pierres angulaires de la recherche en FLE [1]. Or la communication passe aussi par l'interprétation de l'émotion [2]. Néanmoins, il y a peu d'études concernant la perception et la production des émotions en enseignement des langues étrangères. Il s'agit pourtant d'une compétence qu'il faut qu'un apprenant d'une langue étrangère acquière afin de réagir de manière appropriée dans des situations de communication quotidienne où l'émotion est présente. Le but de cette communication est de montrer l'importance de la prosodie dans la détection des émotions dans la parole spontanée. Cette étude fait partie d'un projet plus large qui consiste à évaluer la compétence perceptuelle d'étudiants de FLE à identifier des émotions. Nous voulons souligner la nécessité d'inclure la prosodie dans l'enseignement des langues. Nous présentons ici l'élaboration d'un test de perception basé sur la parole spontanée en français, et qui montre toute l'importance de la prosodie dans la perception de l'émotion.

2. PROTOCOLE EXPÉRIMENTAL

Le but de notre étude est de montrer que la prosodie contient suffisamment d'informations pour renseigner sur l'état émotionnel du locuteur. Nous avons émis l'hypothèse que la prosodie était un des paramètres de l'expressivité des émotions. Nous avons aussi voulu évaluer la capacité d'apprenants d'une langue étrangère à percevoir une émotion avec pour seule information la prosodie de la langue. Pour finir, nous avons émis l'hypothèse que les locuteurs de différents groupes de langue perçoivent différemment la colère.

2.1. Description du corpus d'origine

Nous avons choisi d'utiliser un corpus de parole spontanée pour construire notre test de perception. Nous voulions travailler sur de l'émotion réelle et non pas jouée. Une série de canulars radiophoniques nous a permis d'extraire un corpus intéressant. Un animateur radio appelle des professionnels ou des institutions. Il joue le rôle d'un client et demande un service qui ne correspond pas à ce que propose le professionnel. Il crée ainsi une situation d'incompréhension et de mésentente, qui entraîne une réaction de colère de la part de la victime. Nous avons collecté 24 canulars, ce qui correspond à 1 heure et 4 minutes de parole. Nous en avons éliminé 9, qui ne contenaient pas suffisamment de colère. Nous avons transcrit orthographiquement les 15 restants à l'aide de Transcriber 4.0. Nous avons ensuite annoté chaque tour de parole de chacun de ces dialogues selon trois catégories : Colère (C) associée à un chiffre de 1 à 5, marquant l'intensité de la colère perçue; Neutre (N) pour décrire un tour de parole dans lequel aucune expressivité particulière n'est perceptible ; Autre Emotion (AE) qui décrit un tour de parole dans lequel le locuteur exprime une émotion autre que la colère. Cette dernière annotation est également associée à une échelle de 1 à 5. Cette annotation préliminaire, basée sur notre expérience de locuteurs natifs du français et de spécialistes en phonétique, nous a permis de faire une première sélection des tours de parole.

2.2. Sélection des phrases (pré-test)

A partir de cette première annotation, nous avons effectué un pré-test perceptif pour déterminer l'état émotionnel perçu dans chacun des tours de parole que nous avions sélectionné. Nous avons demandé à 5 sujets francophones natifs d'écouter 81 tours de parole. Ils

devaient juger si l'état émotionnel contenu dans les énoncés qu'ils entendaient était de la colère, une autre émotion ou neutre. En fonction des jugements obtenus à ce pré-test : nous avons choisi 13 énoncés clairement perçus comme étant de la colère à au moins 80% et 13 autres comme ne contenant pas de la colère.

3. Test de perception

3.1. Stimuli

Pour montrer que la prosodie était suffisante pour percevoir la colère contenue dans nos énoncés, nous avons décidé d'isoler la prosodie du contenu segmental. Celui-ci est ainsi neutralisé, de manière à ce qu'il n'interfère pas dans la perception de l'émotion. Pour ce faire, nous avons décidé de masquer le contenu segmental avec du bruit blanc [3]. Ce n'est sans doute pas la méthode de masquage la plus couramment employée, mais elle proposait une alternative plus acceptable que les autres à nos yeux. L'une de ces méthodes consiste à re-synthétiser les énoncés. Nous étions toutefois très attachées à garder l'aspect spontané de notre corpus et nous ne voulions pas le modifier en le re-synthétisant. L'autre méthode consiste en un filtre passe-bas. C'est une bonne méthode de masquage. Mais elle coupe l'énergie dans les hautes fréquences. Or, l'énergie dans les hautes fréquences est un des paramètres acoustiques de la colère [4]. Et il nous paraissait important de conserver ce paramètre dans un premier temps. Nous avons donc préféré la méthode de masquage avec le bruit blanc. Pour chaque énoncé que nous avions choisi à partir du pré-test, nous avons créé un bruit blanc de même durée avec Soundforge 7.0, que nous avons mixé au son d'origine. Nous avons réglé l'intensité du bruit blanc de manière à ce qu'il cache l'information linguistique sans masquer pour autant la voix du locuteur. Ces stimuli ont été doublés, c'est-à-dire présentés deux fois aux sujets, afin de vérifier pour chaque sujet la cohérence des réponses. Nous avons également rajouté 5 stimuli d'entraînement, afin que les sujets puissent s'accoutumer à la sonorité particulière des stimuli, perçus comme des sons de « mauvaise qualité ».

3.2. Sujets

Nous avons demandé à 4 groupes de locuteurs d'effectuer le test de perception : 2 groupes de francophones natifs, et 2 groupes d'apprenants en Français Langue Etrangère.

Le groupe contrôle 1 (Français - sans bruit) est constitué de locuteurs francophones natifs (6 femmes et 4 hommes) ayant passé le test avec les sons originaux, sans addition de bruit blanc. Il s'agissait d'établir les réponses correctes au test. Le groupe contrôle 2 (Français - avec bruit) est constitué de 10 locuteurs francophones (6 femmes et 4 hommes) ayant passé le test avec des stimuli masqués par le bruit.

Le premier groupe testé est composé de 7 femmes et 3 hommes locuteurs du portugais, étudiants à l'Université de Lisbonne, le second de 10 locuteurs tchèques (8 femmes et 2 hommes), étudiants à l'Université de Pilsen. Les deux groupes d'apprenants ont été choisis par les enseignants en fonction de leur niveau, B2 selon le Portfolio Européen des Langues [5]. Aucun n'était déjà venu en France.

3.3. Tâche

Nous avons demandé aux auditeurs d'effectuer une double tâche : une tâche de décision et une tâche d'évaluation.

La première tâche consistait à décider si le son entendu véhiculait de la colère ou non. Si oui, il s'agissait ensuite de déterminer le degré de colère sur une échelle de 1 (faible) à 5 (fort).

Les auditeurs étaient préalablement informés de la mauvaise qualité des sons entendus, et ce afin d'éviter une adaptation trop longue à ces stimuli.

3.4. Interface

Le test de perception a été présenté sous format électronique à l'aide du logiciel EasyPhP. Nous y avons attaché une importance particulière afin de contrôler au maximum l'impact de la forme prise par le test sur les réponses des auditeurs.

Le format électronique nous a notamment permis de contrôler toutes les étapes effectuées par les auditeurs : le temps passé sur chaque décision, les hésitations, sont mesurés, les retours en arrière (incontrôlables avec un test papier) sont impossibles, les auditeurs doivent donner une réponse pour chaque son avant de passer au suivant.

Par ailleurs, nous avons veillé à ce que le test soit pertinent du point de vue de la didactique des langues étrangères. Ainsi, nous avons présenté les consignes dans la langue maternelle (LM) des auditeurs de sorte que celles-ci soient bien comprises. Les stimuli proposés étant en français, nous avons choisi de ne pas utiliser la LM des auditeurs pendant la phase de test. Nous leur avons donc proposé de donner leurs réponses avec des icônes afin d'éviter toute surcharge cognitive due aux multiples changements de code. L'icône « écouter le son » disparaissait après deux écoutes successives (ou après la décision), les icônes « Colère » et « pas colère » apparaissaient au moment d'effectuer le choix concernant l'émotion du stimulus. Pour finir, l'échelle de 1 à 5 était présentée en écriture numérique. Cette interface, relativement flexible, devrait nous permettre d'effectuer d'autres tests avec d'autres émotions et d'autres groupes de locuteurs.

Figure 1 : Exemple d'une étape du test

4. Résultats

Nous allons présenter dans cette section les résultats obtenus jusqu'ici. Pour commencer, nous avons vérifié la cohérence entre les réponses données à la première et à la seconde écoute des stimuli en effectuant un test de corrélation de Spearman (ordonné par rang) avec le logiciel Statview 5.0. La table 1 montre que cette corrélation est significative, c'est-à-dire que les réponses des 4 groupes sont cohérentes pour les deux écoutes.

Table 1 : Résultats tests de corrélation de Spearman.

Français – sans bruit	$\rho_9 = 0.763$	$p < 0.0001$
Français – avec bruit	$\rho_9 = 0.848$	$p < 0.0001$
Portugais FLE	$\rho_9 = 0.833$	$p < 0.0001$
Tchèques FLE	$\rho_9 = 0.691$	$p < 0.0001$

Nous avons ensuite comparé les réponses des auditeurs ayant passé le test bruité avec les réponses des francophones dans le cas des sons non bruités. Ce test nous a permis de montrer que les 3 groupes reconnaissent la colère malgré le bruit : les 13 phrases de colère ont été reconnues comme telles à 84 % par les Français et les Tchèques, et à 77% par les Portugais. On peut donc conclure que la prosodie seule permet de reconnaître cette émotion. Dans un second temps, nous avons examiné la catégorisation de la colère en degrés par les locuteurs.

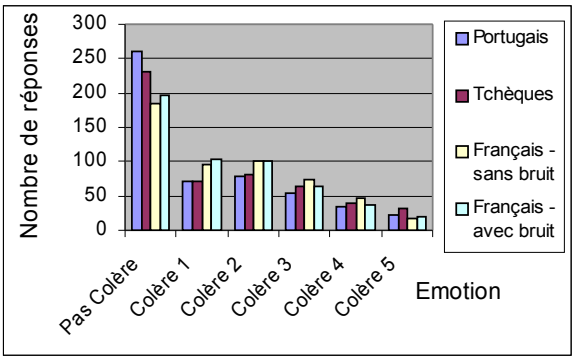


Figure 2 : Nombre et type de jugement par groupe d'auditeurs

Nous pouvons observer dans la figure 2 que les auditeurs ont répondu « Pas colère » dans une plus grande proportion que les autres réponses, ce qui est conforme à ce qui était attendu, la moitié des stimuli proposés ne véhiculant pas de colère. En revanche, on peut noter une différence de l'ordre de 10 % entre les réponses des francophones et celles des apprenants de FLE, différence répercutée dans les catégories « Colère 1 » et « Colère 2 ». Ceci semble indiquer une différence de perception du degré de colère. Tandis que les francophones natifs vont disperser leurs réponses de manière plus homogène entre les différentes catégories de colère faible - Colère 1 (C1) et Colère 2 (C2) - les auditeurs non natifs (ici Tchèques et Portugais) vont percevoir la faible colère comme n'étant pas du tout de la colère. De la même façon, les auditeurs tchèques ont émis un jugement de forte colère (C4 – C5) plus souvent que les Français et les Portugais. Un test statistique sur l'ensemble des résultats (test de Friedman) indique qu'il n'y a pas de différence suffisamment grande pour être significative entre les réponses globales des différents groupes. Cela confirme notre hypothèse première : la prosodie semble suffisante d'après ce test de perception, pour percevoir la colère. Cependant, des tests plus ponctuels (tests t) nous indiquent qu'il y a parfois une différence de perception entre les groupes. Ainsi, la différence entre les réponses des Tchèques et des francophones (test avec bruit) pour la Colère 1 est significative, p=0.0207. Ceci corrobore notre interprétation précédente, les Tchèques ont une perception plus approximative du degré de colère que les francophones natifs. La différence entre les Français (test sans bruit) et les Portugais pour la Colère 3 et la Colère 4 est également significative (respectivement p=0.0338 et p=0,0301). Ce résultat nous indique que les informations sémantiques ont probablement influencé les locuteurs francophones lors du test non bruité pour ces catégories. Des phrases qui ont été jugées comme Colère 3 et 4 dans les conditions de test non bruité ont été catégorisées comme Colère 1 ou 2 lorsque les auditeurs n'avaient pas accès au sens. On peut supposer que lorsque l'information émotionnelle est suffisamment marquée sémantiquement, la prosodie est quant à elle moins marquée car elle serait alors redondante.

5. Analyse prosodique

Afin de déterminer quels étaient les paramètres prosodiques qui pourraient servir d'indices pour la perception de la colère, nous avons procédé à quelques mesures globales, notamment sur la F0 et sur l'intensité. Nous ne présenterons ici que des résultats préliminaires. Une analyse plus affinée est en cours.

Nous avons mesuré pour chacun de nos énoncés la moyenne de la F0, les valeurs minimale et maximale, l'amplitude de la F0. Nous avons fait les moyennes de ces mesures par type de locuteur, homme ou femme, en opposant d'une part les énoncés reconnus comme étant de la colère et d'autre part ceux reconnus comme n'en

étant pas. Si l'on compare les productions des hommes pour les deux catégories, on s'aperçoit qu'au niveau de la F0, il y a une réelle différence. Les énoncés reconnus comme « colère » présentent une moyenne et une amplitude de F0 plus élevées que les énoncés perçus comme « pas colère ». Au niveau de la F0 moyenne, il y a une différence de 40 Hz (environ 4 demi-tons) et pour l'amplitude la différence s'élève à presque 60 Hz (soit trois demi-tons et un quart). En revanche, pour les femmes, il n'y a aucune différence au niveau de la F0 entre les énoncés « colère » et les énoncés « pas colère ». Nous pouvons cependant expliquer ce résultat par le fait que nous avons moins de locutrices que de locuteurs dans notre corpus. Il est donc plus difficile d'établir un résultat certain.

Ces mesures préliminaires globales nécessitent une recherche plus approfondie. Il serait notamment intéressant de voir les différences de registres entre chaque énoncé de « colère » et le registre moyen du locuteur qui prononce cet énoncé. Il nous paraît en effet difficile d'obtenir des résultats stables, en isolant complètement l'énoncé produit du contexte du dialogue. Cette analyse est en cours. Toutefois, nous avons déjà quelques exemples intéressants de courbes mélodiques qui correspondent bien aux patrons de la colère.

Figure 3 : Courbe mélodique d'un énoncé reconnu comme « colère » (Winpitch Pro)

La figure ci-dessus montre la courbe mélodique obtenue avec Winpitch Pro d'un énoncé, prononcé par un locuteur masculin, reconnu comme étant de la colère forte (C4-C5) à 84%. En observant cette courbe, on remarque un contour de F0 très abrupt, avec des montées et des descentes brutales [6]. L'amplitude de F0 sur certaines syllabes est de l'ordre de 100 Hz. La variation de F0 est donc très importante. Le registre de la voix du locuteur pour cet énoncé est très étendu : il passe de 88Hz à 318 Hz (22 demi-tons de différence) dans le même énoncé. Si d'un point de vue global les mesures prises pour l'instant ne sont pas très significatives, en s'intéressant à une analyse plus fine de chaque énoncé, de véritables différences devraient apparaître, notamment au niveau de la courbe de F0.

6. Conclusion et perspectives

Cette étude nous a permis de montrer que la prosodie jouait un rôle important dans la perception de la colère en français spontané. Elle a mis également en avant la capacité d'apprenants de FLE à percevoir l'émotion dans la langue étrangère.

Cette étude n'est que la première partie d'un projet plus global. Nous souhaiterions dans un second temps, travailler sur la reproduction des émotions par les apprenants de FLE, en essayant de voir les interférences de la langue maternelle dans leurs réalisations. D'autre part, nous souhaitons mener quelques études de perception secondaires : nous allons tester par exemple les différentes méthodes de masquage du son, pour déterminer quelle est la plus adaptée à notre étude. Un test est en cours avec un filtre passe-bas.

Ce test a été créé dans l'idée d'être adaptable à des apprenants de différentes langues maternelles, et à des émotions différentes, et nous aimerions dans un proche avenir pouvoir le développer.

Bibliographie

[1] J. Beale, Is communicative language teaching a thing of the past ?, Babel, Vol. 37, No. 1, Winter 2002, pp. 12-6.

[2] E. Galazzi et E. Guimbretière, Intonation et attitudes: une question de perception, Studi di Linguistica, Storia della lingua Filologia francesi, Edizioni dell'Orso, Milan, 1994.

[3] Miller and Licklider, The intelligibility of interrupted speech, J. Acoust. Soc. Am. 22: 167-173, 1950)

[4] T. Bänziger, and K. R. Scherer, Relations entre caractéristiques vocales perçues et émotions attribuées. Actes des Journées Prosodie 2001, Grenoble, France, 119-124, 2003.

[5] http://www.enpc.fr/fr/international/eleves_etrangers/portfolio. pdf

[6] P. Léon, Précis de phonostylistique : Parole et expressivité, Nathan, Paris, France, 1993.

La production et la perception des voyelles orales françaises par les apprenants japonophones

KAMIYAMA Takeki

Laboratoire de phonétique et phonologie (UMR 7018) CNRS / Sorbonne Nouvelle - Paris III
19, rue des Bernardins, 75 005 Paris, France
takekik@phiz.c.u-tokyo.ac.jp
http://www.cavi.univ-paris3.fr/ilpga/ed/student/stkt/

ABSTRACT

In order to examine the production and perception of French oral vowels by native speakers of Japanese learning French as a foreign language, a series of experiments were conducted. First, 10 isolated oral vowels pronounced by 4 native speakers of French (2 male and 2 female) were identified by 5 Japanese-speaking learners. Second, the formant frequencies were measured for the vowels that were 1) read, and 2) repeated after native speakers' recordings, by 3 learners. The results suggest that it is difficult to perceive and produce in a native-like manner not only "new" vowels (front rounded series) but also a "similar" one (such as the French high back vowel /u/: Flege [3]), as well as open-mid / close-mid oppositions.

1. INTRODUCTION

Les phonéticiens et les psycholinguistes se sont intéressés à la production et la perception des phonèmes dans les langues non-natives du locuteur.

Lambacher, Martens et al. [6] ont montré que les apprenants japonais ont de la difficulté pour identifier les voyelles postérieures /ʌ/ /ɔ/, et les voyelles ouvertes /æ/ et /ɑ/ de l'anglais américain. Strange, Akahane et al. [9] ont conclu que les voyelles extrêmes (/i/ /ɑ/ /u/) de l'anglais américain sont assimilées aux trois voyelles correspondantes du japonais (/i/ /a/ /u/) d'une façon relativement stable, mais moins pour les autres, et que les voyelles longues (et diphtongues) ont été perçues comme similaires aux voyelles longues du japonais, mais seulement dans des phrases cadre (et pas dans des mots isolés). Gottfried [5] a montré la difficulté que les auditeurs américains (anglophones) rencontrent face aux voyelles antérieures arrondies, qui n'existent pas dans le système phonémique de l'anglais. L'étude de Strange, Levy et al. [10] sur l'assimilation perceptive des voyelles françaises par les américains suggère que la série antérieure arrondie est perçue plus similaire aux voyelles postérieures qu'antérieures. Flege [3] a montré que la voyelle /y/ prononcée par des locuteurs natifs de l'anglais américain et expérimentés en français n'est pas différente de celle prononcée par les natifs français, mais le F2 du /u/ français prononcé par tous les locuteurs américains est significativement plus élevé que celui des Français natifs. Ceci suggère qu'il est plus difficile de produire d'une manière « authentique » (comme les natifs) les sons similaires mais différents par rapport aux équivalents dans le système phonémique du L1 (ex. /u/), que les sons nouveaux, qui n'ont pas d'équivalent dans le système du L1 (ex. /y/).

Cependant, pour ce qui est de la production et la perception des voyelles orales du français par les apprenants japonophones, nous ne connaissons pas d'études expérimentales effectuées sur ce sujet.

Le système vocalique du japonais est composé de 5 voyelles /i/ /e/ /a/ /o/ /u/. Nous pouvons donc considérer les voyelles antérieures arrondies du français comme des sons « nouveaux », et l'opposition entre les mi-fermées et les mi-ouvertes comme une opposition « nouvelle ». En revanche, le /u/ serait considéré comme un son « similaire », car l'équivalent en japonais présente un F2 supérieur à 1000 Hz (Figure 1, Sugito [11]), ce qui n'est pas le cas pour le /u/ français. Du point de vue articulatoire, cette voyelle, transcrite souvent [ɯ], est moins arrondie est plus antérieure que le [u] à la française, comme le montrent les profils articulatoires de Uemura [12].

Nous pouvons postuler l'hypothèse suivante : les voyelles antérieures arrondies du français ainsi que l'opposition entre les mi-ouvertes et les mi-fermées seraient difficiles à percevoir et à produire pour les apprenants japonophones (au moins pour ceux qui sont au niveau élémentaire), et ce serait encore plus le cas concernant la voyelle /u/. Les voyelles « similaires » sont-elles plus difficiles à acquérir en production et en perception que les voyelles « nouvelles » ? Afin de vérifier cette hypothèse, une série d'expériences de production et de perception a été effectuée.

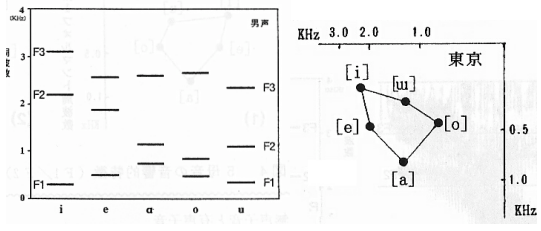


Figure 1 : Les trois premiers formants (à gauche) et les deux premiers formants en 2 dimensions (à droite ; F1 sur l'axe vertical et F2 sur l'axe horizontal) des 5 voyelles (voix d'homme) du japonais de Tokyo (Sugito [11]).

2. Les voyelles cible (voyelles orales du français)

2.1. Études antérieures

Concernant les valeurs formantiques des voyelles orales du français, nous disposons des données de CALLIOPE [2], ainsi que celles de Gendrot & Adda [4] recueillies sur un grand corpus radiophonique. Nous observons, entre autres, les caractéristiques suivantes : **1)** F3 élevé, F3 et F4 proches pour /i/, **2)** F1 et F2 proches et très bas pour /u/ (moins vrai en [4]), **3)** F2 et F3 proches aux environs de 1800 (hommes) / 2300 (femmes) Hz pour /y/, **4)** F2 moyen (1300-1500 Hz pour les hommes, 1600-1700 Hz pour les femmes) pour /ø/ et /œ/.

2.2. Les données de la présente étude

4 locuteurs natifs du français dit standard ont prononcé les 10 voyelles orales françaises dans une phrase cadre (« Je dis /V/ comme dans … »). Pour chacune des voyelles cible, les 4 premiers formants sur 5 points (1. du début jusqu'à un cinquième de la durée de la voyelle, 2. le deuxième cinquième, … 5. jusqu'à la fin) ont été mesurés sur Praat [1]. Parmi ces 5 valeurs, celles qui sont discontinues et aberrantes ont été éliminées après vérification du spectrogramme et/ou du spectre. Comme attendu des voyelles du français, aucune n'était diphtonguée. La table 3 indique la valeur moyenne de 3 répétitions de chaque locuteur. Nous pouvons retrouver les tendances décrites dans la section 2.1.

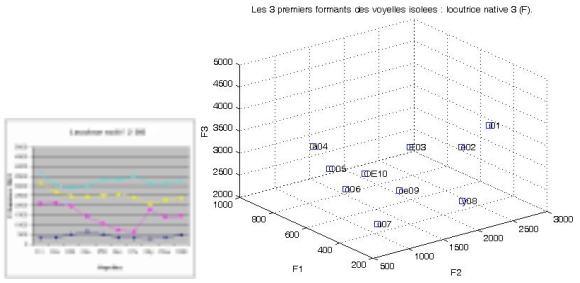
Table 1 : Les 4 premiers formants des voyelles orales du français en Hertz (moyenne de 3 répétitions).

Loc.1 (H)	F1	F2	F3	F4
i	323	2194	3140	3829
e	441	2048	2576	3479
ɛ	512	1891	2556	3569
a	699	1261	2392	3434
ɔ	495	951	2629	3401
o	372	688	2618	3496
u	289	631	2489	3503
y	284	1806	2100	3290
ø	396	1276	2304	3271
œ	463	1366	2390	3386

Loc.2 (H)	F1	F2	F3	F4
i	312	2091	3160	3679
e	332	2123	2691	3041
ɛ	505	1964	2478	2879
a	654	1396	2444	2993
ɔ	505	1053	2501	3336
o	330	715	2565	3317
u	304	637	2416	3514
y	288	1749	2089	3128
ø	324	1372	2282	3144
œ	479	1461	2401	3251

Loc.3 (F)	F1	F2	F3	F4
i	347	2519	3903	4439
e	500	2488	3162	3985
ɛ	686	2190	2972	3917
a	869	1238	3071	3977
ɔ	694	1051	2940	4078
o	481	776	2968	4024
u	302	798	2462	3435
y	306	2032	2471	3804
ø	490	1570	2583	4012
œ	669	1733	2778	4149

Loc.4 (F)	F1	F2	F3	F4
i	343	2451	3536	3967
e	558	2321	2966	3914
ɛ	696	2088	2667	3562
a	845	1441	2612	3731
ɔ	735	1108	2826	3827
o	515	912	2835	3878
u	385	744	2311	3457
y	378	1989	2522	3733
ø	556	1431	2583	3954
œ	689	1551	2635	3987



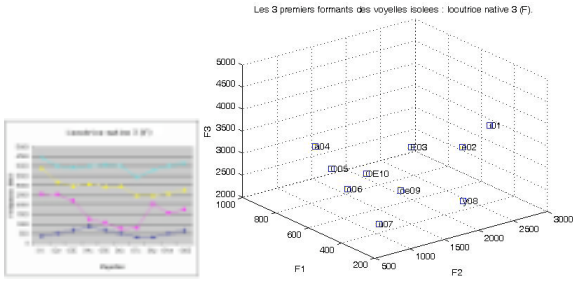


Figure 2 : La représentation graphique des valeurs de la table 3. Locuteur 2 en haut, et locutrice 3 en bas.

3. La perception des voyelles orales du français par les apprenants japonophones

Les voyelles qui ont été présentées dans la section précédente ont été utilisées comme stimuli d'un test d'identification effectuée auprès de 5 auditeurs natifs du japonais apprenant le français langue étrangère dans un contexte exolingue (à l'Université de Tokyo, Japon). Leur expérience d'apprentissage variait de 3 mois à 2 ans, mais tous avaient appris le français au Japon, et par conséquent ils n'avaient pas eu beaucoup d'input auditif dans leur vie quotidienne.

Ils ont écouté des voyelles isolées (4 locuteurs * 2 répétitions * 13 voyelles, y compris les 3 voyelles nasales) et choisi la voyelle qu'ils pensaient avoir perçue.

Le résultat présenté dans le tableau 4 montre les faits suivants : **1)** comme attendu, les mi-ouvertes et les mi-fermées sont largement confondues ; **2)** cette tendance est très marquée pour /e/-/ɛ/, mais le résultat est plus compliqué pour les autres : /ɔ/ a été perçu comme /ø/ ou /œ/ (2+3/40), /o/ comme /u/ (8/40), /ø/ comme /u/ (12/40) ; **3)** il y a une confusion importante entre /u/ et /ø/ dans les deux sens, et une tendance similaire mais moins marquée entre /y/ et /ø/.

Table 2 : Les résultats du test d'identification (8 stimuli par voyelle * 5 auditeurs).

stimuli\réponse	i	e	ɛ	a	ɔ	o	u	y	ø	œ	ɛ̃	ɑ̃	ɔ̃	total_stimuli
i	39	1	0	0	0	0	0	0	0	0	0	0	0	40
e	5	16	16	0	0	0	0	0	1	1	1	0	0	40
ɛ	0	15	24	0	0	0	0	0	0	1	0	0	0	40
a	0	0	0	34	0	0	0	0	1	2	0	3	0	40
ɔ	0	0	0	1	24	6	0	0	2	3	1	3	0	40
o	0	0	0	0	6	23	8	0	3	0	0	0	0	40
u	0	0	0	0	0	0	22	2	12	0	0	0	4	40
y	0	0	0	0	0	0	0	33	6	1	0	0	0	40
ø	0	0	0	0	0	0	12	5	17	6	0	0	0	40
œ	0	0	0	0	0	0	1	1	11	24	1	2	0	40
ɛ̃	0	0	0	5	0	0	0	0	0	2	21	12	0	40
ɑ̃	0	0	0	0	8	2	0	0	0	0	3	14	13	40
ɔ̃	0	0	0	0	3	5	0	0	3	1	5	8	15	40

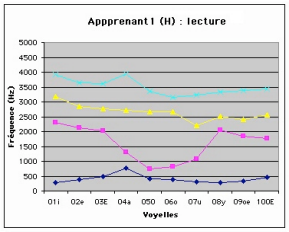
4. La production des voyelles orales du français par les apprenants japonophones

4.1. Lecture

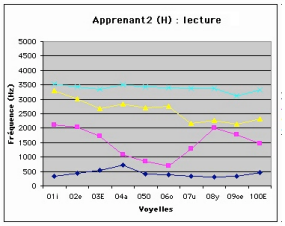
3 apprenants (2 hommes et 1 femme) ont lu les voyelles

orales du français dans la même phrase cadre que celle de la section 2.2. Les phrases ont été répétées 2 fois. Nous observons les tendances suivantes (figure 3) : **1)** les apprenants 1 et 2 ont prononcé le /y/ avec ses F2 et F3 proches (malgré un F2 supérieur à 2000 Hz), tandis que l'apprenante 3 a produit une diphtongue ouvrante (figures 4 et 5), qui ressemble à la syllabe /ju/ du japonais ; **2)** tous les trois apprenants ont prononcé la voyelle « similaire » /u/ avec un F2 plus élevé que celui de /o/ et de /ɔ/ (notons que les apprenants produisent ces deux voyelles avec une valeur de F2 qui n'est pas bien différente des locuteurs natifs), ce qui rapproche cette voyelle au /ø/ (et peut-être au [ʊ] du français québécois [7]). De plus, l'intensité relative des formants supérieurs est plus importante que dans le /u/ des locuteurs français, ce qui fait qu'il y a plus d'énergie sur les moyennes et les hautes fréquences (figure 6). **3)** Comme attendu, les différences entre les mi-fermées et les mi-ouvertes sont moins bien marquées que chez les locuteurs natifs.

Appr.1 (H)	F1	F2	F3	F4
i	285	2297	3188	3923
e	382	2135	2855	3643
ɛ	486	2015	2770	3606
a	772	1304	2709	3955
ɔ	407	749	2656	3360
o	376	830	2665	3164
u	316	1065	2212	3234
y	281	2062	2523	3346
ø	334	1850	2414	3383
œ	465	1770	2555	3433



Appr.2 (H)	F1	F2	F3	F4
i	324	2114	3290	3523
e	444	2045	3006	3420
ɛ	543	1729	2680	3361
a	721	1072	2843	3502
ɔ	414	852	2703	3434
o	391	683	2751	3403
u	326	1283	2168	3384
y	317	2005	2256	3375
ø	336	1767	2144	3126
œ	468	1474	2318	3332



Appr.3 (F)	F1	F2	F3	F4
i	411	3101	3865	4647
e	551	2547	3405	4621
ɛ	553	2693	3471	4653
a	785	1452	3175	4426
ɔ	610	1027	3380	4255
o	558	888	3282	4172
u	406	1388	3250	4514
y	399	1945	3018	4257
ø	504	1429	3146	4276
œ	538	1382	3068	4181

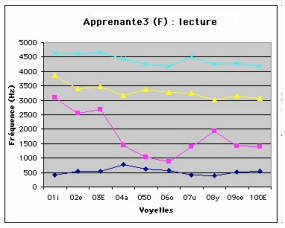


Figure 3 : Les 4 premiers formants des voyelles orales du français lues par 3 apprenants japonais (moyenne de 2 répétitions).

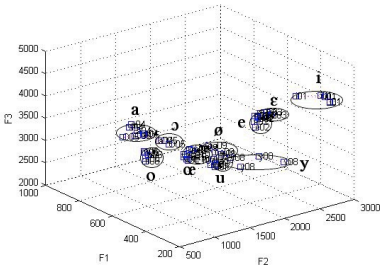


Figure 4 : Les 3 premiers formants de l'apprenante 3 dans le temps (5 points représentés).

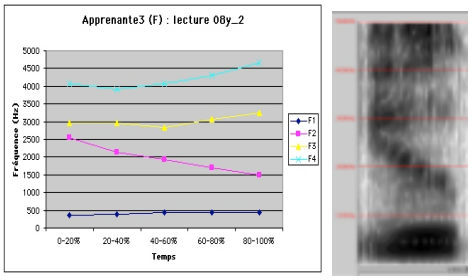


Figure 5 : Le changement des valeurs formantiques et le spectrogramme du /y/ lu par l'apprenante 3.

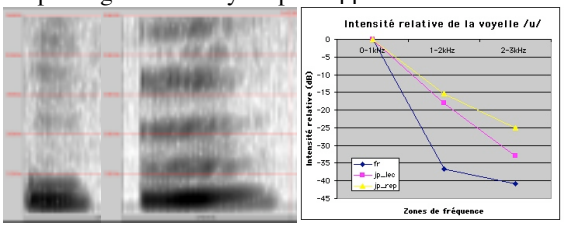


Figure 6 : Les spectrogrammes de la voyelle /u/ prononcée par le locuteur natif 1 (à gauche) et par l'apprenant 2 (au milieu). L'intensité relative sur 3 zones de fréquence (0-1 kHz, 1-2 kHz, 2-3 kHz) du /u/ prononcé par les locuteurs natifs (12 occurrences : fr), lu (6 : jp_lec) et répété par les apprenants (10 : jp_rep).

4.2. Répétition

Les mêmes apprenants ont répété les voyelles isolées après le « modèle » prononcé par les locuteurs natifs (2 hommes pour les deux apprenants, 2 femmes pour l'apprenante).

Nous observons les mêmes tendances que la lecture (figure 7 pour l'apprenant 1), mais le F2 du /u/ est encore plus élevé qu'en lecture chez deux apprenants. Ceci indiquerait la difficulté liée à la perception de cette voyelle.

L'apprenante 3, qui a diphtongué le /y/ en lecture (comme le /ju/ du japonais) a fait de même après une voyelle relativement brève (133 ms : locutrice 4), mais pas après une voyelle longue (286 ms : locutrice 3), malgré un écart trop important entre F2 et F3 (1730 Hz et 2876 Hz) pour un "beau" /y/ français. Il est possible que la longue durée ait permis à l'apprenante de se rendre compte que le timbre de la voyelle était nettement différent de celui de la deuxième partie de /ju/ japonais, lui permettant de percevoir que ce n'était pas une diphtongue ouvrante, et ainsi d'éviter de produire la séquence /ju/ à la japonaise.

Appr.1 (H)	F1	F2	F3	F4
i	262	2294	3003	3768
e	385	2011	2762	3579
ɛ	472	1939	2721	3707
a	699	1171	2605	4072
ɔ	427	890	2625	3090
o	362	823	2444	3022
u	313	1362	2185	3171
y	268	1954	2434	3342
ø	343	1527	2329	3280
œ	423	1341	2529	3223

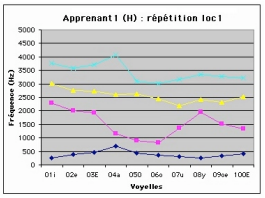


Figure 7 : Les valeurs formantiques des voyelles orales du français répétées par l'apprenant 1.

5. CONCLUSION

Les résultats de la présente série d'expérience suggèrent que les apprenants japonophones auraient effectivement des difficultés pour apprendre à percevoir et à produire 1) la nouvelle opposition mi-fermées / mi-ouvertes, 2) les nouvelles voyelles (antérieures arrondies), et 3) la voyelle similaire /u/.

Il reste à vérifier la difficulté relative de la voyelle similaire /u/ (aucun apprenant ne l'a prononcée comme les natifs), et des nouvelles voyelles (/y/ prononcé avec F2 et F3 proches par 2 apprenants sur 3), sur un plus grand nombre d'apprenants de même niveau, et de différents niveaux afin d'examiner l'ordre d'acquisition.

D'ailleurs, le statut de /y/, que nous avons classé comme nouvelle voyelle, est intéressant à discuter. Cette voyelle est adaptée en japonais systématiquement, quel que soit le contexte consonantique, à la séquence /ju/, ainsi que la séquence /jø/ du français : Hugo /jugoː/, Camus /kamju/, Dumas /djuma/, Curie /kjuriː/, rue /rjuː/), comme l'a fait l'apprenante 3. Ceci nous permettrait de classer cette voyelle entre les deux catégories « nouvelle » et « similaire ». Il sera intéressant de tester des connaissances méta-phonologiques des apprenants.

Ce qui est intéressant est la confusion de perception entre /u/ et /ø/ (table 4), à la différence des auditeurs américains (similarité entre /y/-/u/, ou /y/-/ø/ [5][6]), car ces deux voyelles ne partagent ni le même degré d'ouverture ni d'antériorité dans le système phonologique du français. Sur le plan acoustique, les formants et leur intensité relative du /u/ prononcé par les apprenants (ex. figure 6) ressemblent effectivement au /ø/, sauf la valeur basse du F1, mais les Français semblent percevoir "une voyelle entre /u/ et /y/", mais pas le /ø/ (à vérifier dans un test de perception en préparation), ce qui ne correspond pas à la perception des apprenants eux-mêmes (très peu de confusions entre /u/ et /y/). Tout cela suggère que le F1 serait plus important à l'oreille française, sans doute à cause des 4 degrés d'ouverture. Ce phénomène reste encore à examiner en utilisant la synthèse articulatoire et à formant.

Afin d'examiner plus profondément la perception et la production des voyelles orales du français, un certain nombre d'autres expériences sont en cours ou en préparation, y compris un test de discrimination auprès des apprenants, un test d'identification et de discrimination des voyelles prononcées par les apprenants. Le présent article traite uniquement des voyelles isolées, mais des analyses et des tests de perception des voyelles en contexte (CVCVCV) seront effectués pour examiner la variabilité selon le contexte.

REMERCIEMENTS

L'auteur remercie sa directrice de recherche Jacqueline Vaissière et Cécile Fougeron, ainsi que les deux relecteurs anonymes pour la relecture d'une version antérieure et leurs commentaires précieux.

BIBLIOGRAPHIE

[1] P. Boersma and D. Weenink. PRAAT, a system for doing phonetics by Computer. *Glot International*, 5(9/10): 341-345, 2001.

[2] CALLIOPE. *La parole et son traitement automatique*. Masson, Paris, Milano, Barcelona, Mexico, 1989.

[3] J. E. Flege. The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics*, 15: 47-65, 1987.

[4] C. Gendrot and M. Adda-Decker. Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande. In *Proc. Colloque MIDL 2004*, pages 7-12, 2004.

[5] T. L. Gottfried. Effects of consonant context on the perception of French vowels. Journal of Phonetics, 12: 91-114, 1984.

[6] S. Lambacher, W. Martens, et al. Comparison of identification of American English vowels by native speakers of Japanese and English. In *Proc. Phonetic Society of Japan 2000*, pages 213-217, 2000.

[7] P. Martin. Le système vocalique du français du Québec. De l'acoustique à la phonologie. *La linguistique*, 38(2) : 71-88, 2002.

[8] W. Strange (Ed.). *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research*. York Press, Timonium, MD, 1995.

[9] W. Strange, R. Akahane-Yamada, et al. Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics*, 26: 311-344, 1998.

[10] W. Strange, E. Levy, et al. Perceptual assimilation of French and German vowels by American English monolinguals: Acoustic similarity does not predict perceptual similarity. *Journal of the Acoustical Society of America*, 115(5): 2606-2606, 2004.

[11] M. Sugito. *Oosaka - Tookyoo akusento onsei jiten CD-ROM: kaisetsuhen* (CD-ROM Accent dictionary of Spoken Osaka and Tokyo Japanese). Maruzen, Tokyo, 1995.

[12] Y. Uemura. *Nihongo no boin, shiin, onsetu: chouon undou no jikken-onseigakuteki kenkyuu* (Vowels, consonants and syllables in Japanese: an exprimental phonetic study on articulatory mouvements), Shuuei shuppan, Tokyo, 1990.

Reconnaissance de la parole guidée par des transcriptions approchées

Benjamin Lecouteux, Georges Linarès, Pascal Nocera, Jean-François Bonastre

Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, B.P 1228, 84911 Avignon, France
{benjamin.lecouteux, georges.linares, pascal.nocera, jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

In many cases, an approximated transcript can be associated to speech signal : movies subtitles, scenario and theatre, summaries and radio broadcast. These transcripts correspond rarely to the exact word utterances. The goal of this work is to use these information to improve the performance of an automatic speech recognition (ASR) system with integration of information resulting from the transcripts. In this paper we use the partial transcript in order both to adapt the language model and to rescore the ASR word hypothesis when the partial transcript matches the input signal. Multiple applications are possible : to help deaf people to follow a play with closed caption aligned to the voice signal (with respect to performer variations), to watch a movie in another language using aligned closed captions, to transcript in real time debates or meetings.

1. INTRODUCTION

Dans certaines situations, des sources d'informations externes au signal de parole à transcrire peuvent être disponibles : émissions comportant des sous-titres ou pour lesquelles un résumé peut être mis à disposition, scénario pour le cinéma, scripts de pièces de théâtre, prompteur d'un journaliste... Ces informations peuvent être exploitées pour améliorer les performances d'un système de reconnaissance automatique de la parole (SRAP).

Ce problème est abordé dans le domaine de l'alignement automatique de textes sur des flux audio. Cependant, dès que les informations disponibles s'éloignent de ce qui est réellement dit, un simple alignement forcé entre le signal audio et le texte devient insuffisant : Il est alors nécessaire de retrouver le contenu du message par une transcription automatique de celui-ci, tout en tirant profit des informations évoquées précédemment.

L'utilisation de textes approchés pour le décodage audio a déjà été étudiée dans le cadre de l'indexation audio/vidéo automatisée ([2]) ou de la ré-estimation de modèles acoustiques ([4],[5]) à partir de données non transcrites. Cette tâche est présentée dans la littérature comme un problème de synchronisation de texte sur le flux audio, ou, plus rarement, comme un problème de correction de transcriptions imparfaites. Dans les deux cas, la principale difficulté de la tâche tient à la qualité, souvent approximative, des transcriptions : Placeway et Lafferty mesurent un taux de différence entre les sous-titres d'un film et la transcription exacte compris entre 10% et 20% [9]. Ces divergences entre la transcription et le message réel augmentent considérablement la difficulté de l'alignement.

Dans ce papier nous présentons tout d'abord les problèmes liés à l'alignement d'un texte approché sur un flux de parole. Puis nous décrivons notre solution pour la reconnaissance de la parole guidée par une transcription approchée. Cette solution est basée sur un décodeur à pile asynchrone permettant l'intégration des informations issues des transcriptions approchées. Enfin, les expériences menées et les résultats obtenus sont détaillés avant de présenter quelques conclusions et perspectives.

2. ÉTAT DE L'ART

2.1. Alignement forcé avec transcription exacte

Le sujet de l'alignement sur transcription exacte est abordé par Moreno et Joerg pour aligner de longs documents audio avec leur transcription dans le cadre d'une indexation automatique de documents multimédias [7]. Moreno et Joerg proposent une méthode basée sur la recherche de zones bien synchronisées, appelées *îlots de confiance* [7]. Dans un premier temps, un modèle de langage est estimé sur la transcription exacte. Une première passe isole des zones avec une forte correspondance entre transcription *a priori* et transcription automatique correspondent. Le document est alors segmenté par ces îlots de confiance ; sur chaque segment, un modèle de langage spécifique est estimé. L'algorithme est lancé récursivement sur chaque partie non alignée jusqu'à convergence. Cette méthode, restreinte aux transcriptions exactes, obtient d'excellents résultats : 99% des mots sont correctement alignés.

2.2. Alignement de transcriptions approchées

Le problème du traitement automatique de transcriptions approchées a été abordé par Placeway et Lafferty qui ont expérimenté l'exploitation de sous-titres avec un décodeur synchrone (SPHINX-3) [9]. Leurs expériences portent sur une base de données de journaux diffusés en anglais. Ils proposent d'utiliser les sous-titres en estimant un modèle de langage sur ces derniers puis en alignant le flux audio sur ces sous-titres.

Pour combiner l'information des sous-titres avec les modèles du décodeur, Placeway et Lafferty ont interpolé un modèle de langage générique avec un modèle estimé sur les sous-titres [9]. Ce modèle interpolé est ensuite utilisé par le SRAP. Leurs expérimentations sont menées avec des sous-titres comportant 9.7% d'erreurs par rapport à la transcription exacte. Cette technique améliore les performances du décodage de 15% relatifs de WER (Word Error Rate) par rapport au décodage initial (de 55.8% à 47.2%).

Par ailleurs, ils proposent d'intégrer un mécanisme d'alignement sur les sous-titres, en plus de l'interpolation des modèles : au fur et à mesure que le décodeur avance et propose sa liste de mots candidats, les mots correspondants à la transcription approchée sont favorisés dans le faisceau d'hypothèses. Cette méthode apporte un gain relatif de 37% WER comparé au décodage initial en ramenant le WER à 35%. Le résultat final reste cependant nettement inférieur à la qualité de la transcription approchée fournie au système (9.7% de WER).

3. RECONNAISSANCE DE LA PAROLE PAR DES TRANSCRIPTIONS APPROCHÉES

Notre objectif est d'exploiter des transcriptions approchées avec un décodeur asynchrone basé sur l'algorithme A*, dans le cadre d'un système de broadcast news en langue française. Nous présentons les particularités de ce type de décodeur et la solution pour intégrer l'information contenue dans les transcriptions imparfaites.

Deux méthodes exploitant la transcription approchée ont été expérimentées. La première consiste à combiner un modèle de langage générique et un modèle de langage estimé sur la transcription approchée. La seconde propose d'intégrer un algorithme d'alignement temporel au sein de l'algorithme A*, en influençant directement la fonction d'estimation de l'agorithme de recherche.

3.1. Adaptation des modèles de langage

Dans notre approche, la variabilité linguistique peut être réduite grâce à l'apport d'une transcription exacte ou approchée du discours. Un gain peut être obtenu en réduisant globalement l'espace linguistique, par estimation d'un modèle de langage sur la transcription elle même. Cependant, ce modèle de langage ne suffit pas lorsque le locuteur s'éloigne de la transcription. Un modèle de langage générique est donc interpolé au modèle de langage réduit issu de la transcription approchée.

3.2. Décodeur asynchrone et algorithme d'alignement

Le LIA a développé un système de reconnaissance de la parole grand vocabulaire et parole continue nommé SPEERAL [8]. Le décodeur de SPEERAL, à pile asynchrone, est basé sur l'algorithme de recherche A^*. L'exploration du graphe (le treillis de phonèmes) est dirigée par une fonction d'estimation basée sur deux informations : le score de l'hypothèse courante et une sonde estimant le coût minimal en fin de chemin. Cette sonde h combine un score acoustique et un score d'anticipation linguistique ([3]). Le terme acoustique est calculé par un décodage acoustico-phonétique réalisé par l'algorithme de Viterbi arrière sur le treillis de phonèmes.

L'algorithme A^* repose complètement sur la fonction d'estimation $F(n)$ qui évalue, pour chaque noeud exploré, le coût minimal des chemins passant par ce point. Cette fonction guide l'exploration du graphe d'hypothèses en orientant le décodage vers les chemins dont l'estimation partielle est bonne. Par ailleurs, la progression dans le graphe affine l'évaluation des chemins explorés, ce qui peut conduire à l'abandon de certaines voies. Dans ce cas, l'algorithme revient en arrière et explore d'autres branches du graphe, ce qui le désynchronise du flux audio.

Afin de pouvoir prendre en compte les informations issues de la transcription approchée, un module a été rajouté pour influencer le score de l'hypothèse courante au sein de l'algorithme de recherche. Ce mécanisme oriente le décodage en influençant dynamiquement le score de l'hypothèse évaluée. L'algorithme proposé se décompose en deux parties : la synchronisation entre la transcription et le flux de parole puis l'intégration de l'information issue de l'alignement synchrone à la fonction d'évaluation $F(n)$.

Synchronisation du flux audio et de la transcription imparfaite

Le moteur de reconnaissance construit des hypothèses au fur et à mesure qu'il avance dans le treillis de phonèmes. Les meilleures hypothèses à un instant t sont prolongées. Les modifications apportées au décodeur permettent d'aligner à la transcription approchée chaque nouveau mot et son historique. Ceci est réalisé par un algorithme d'alignement temporel (Dynamic Time Warping [1]). Ainsi, tous les mots de l'hypothèse courante du décodeur sont alignés sur un passage de la transcription. Chaque mot rajouté dans l'hypothèse est alors rescoré en fonction d'un indice de confiance issu de l'alignement.

La figure 1 présente l'évolution des hypothèses du décodeur à pile A* influencées par un alignement DTW sur une transcription approchée.

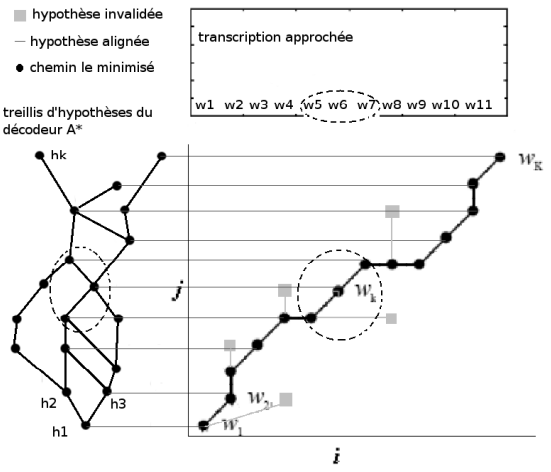


FIG. 1: Synchronisation du faisceau de recherche avec la transcription imparfaite par algorithme DTW

Pondération de l'hypothèse courante en fonction de l'alignement

La fonction d'estimation calcule, pour chaque noeud du graphe, les coûts du chemin exploré ainsi qu'une sonde minimisant le coût des chemins finaux. La qualité de cette sonde influence directement les performances de l'algorithme de recherche. La solution proposée réhausse le score des mots présents dans le faisceau de recherche lorsqu'ils sont alignés avec la transcription approchée ; s'ils ne sont pas présents dans le faisceau, l'algorithme d'alignement n'intervient pas. Pour que notre alignement oriente le moteur de reconnaissance, il faut que l'algorithme A* pré-

sente le mot : le score de l'hypothèse courante sera alors modifié en conséquence. Nous ne modifions pas les scores d'anticipation linguistique.

Une fois l'hypothèse synchronisée avec la transcription, l'algorithme estime un score de synchronie locale, calculé à partir du nombre de mots de l'historique correctement alignés à la transcription. Maximal lorsque le trigramme complet est aligné, ce score décroît en fonction des défauts d'alignement de l'historique.

4. EXPÉRIMENTATIONS

4.1. Cadre expérimental

L'ensemble des expériences a été effectué avec le système de "Broadcast news" développé au LIA qui a été engagé dans la campagne d'évaluation ESTER ([6]).

Corpus utilisé et transcription approchée

Le système est évalué sur une heure d'émission radio issu du corpus de développement d'ESTER (France Inter du 08-04-2003, de 7h à 8h). Ensuite, 10% d'erreurs ont été introduites manuellement dans la transcription, tout en prenant soin de garder une forme journalistique correcte pour respecter le style classique d'une émission radiophonique. Nous simulons ainsi une transcription imparfaite proche de ce que serait le script d'une émission de ce type.

Interpolation de modèles de langages

Des expériences préliminaires ont été menées afin d'identifier les gains potentiels de nos méthodes. Un modèle de langage a été appris sur la transcription exacte, puis combiné avec un modèle de langage générique (65000 mots appris sur Le Monde). L'objectif est d'être capable de mesurer l'effet réel des techniques proposées sur les performances du décodeur. Les mots hors vocabulaire ont été extraits de la transcription pour être phonétisés et ajoutés au modèle de langage. Le tableau 1 présente les résultats d'interpolation d'un modèle de langage appris sur la transcription exacte avec le modèle de langage générique. Par ailleurs, un décodage normal avec modèle générique présente un WER de 22.7%.

TAB. 1: Résultats des expériences de référence avec interpolation du modèle de langage générique (ML-G) et du modèle appris sur la transcription exacte (ML-TrExact)

	Taux d'erreur
ML-G seul	22.7%
ML-TrEx seul	5.2%
ML-G 70% + ML-TrExact 30%	13.0%
ML-G 50% + ML-TrExact 50%	11.5%
ML-G 30% + ML-TrExact 70%	10.8%

Ensuite, à partir de la transcription approchée, un modèle de langage a également été généré. Les expériences utilisant ce modèle de langage combiné au modèle de langage générique sont présentées dans le tableau 2.

Ces expériences montrent qu'un décodage restreint avec un modèle de langage appris sur une transcription approchée améliore sensiblement le taux d'erreurs. Cependant, sans autre source d'information, le moteur de reconnaissance continue à faire des erreurs sur les parties qui ont été mal apprises (les erreurs dans la transcription). Par ailleurs, un décodage avec un modèle de langage appris sur la transcription exacte montre que cette technique ne permet pas de descendre en deçà de 15% de WER.

TAB. 2: Résultats des expériences d'interpolation de modèles de langage Générique (ML-G) et appris sur la transcription approchée (ML-TrErr)

	Taux d'erreur
ML-TrErr seul	16.3%
ML-G 70% + ML-TrErr 30%	16.2%
ML-G 50% + ML-TrErr 50%	15.4%
ML-G 30% + ML-TrErr 70%	15.2%

Expériences avec interpolation de modèles et alignement

Après avoir expérimenté des modèles interpolés, nous avons évalué la méthode de décodage guidée par la transcription qui a été décrite précédemment. Bien que cette approche puisse permettre de lever certaines des limites observées dans la combinaison de modèles, des sources potentielles d'erreur subsistent. En particulier, des heuristiques sont utilisées dans le décodeur pour réduire l'espace de recherche et accélérer le décodage. En conditions normales, ces coupures ne doivent introduire que peu d'erreurs ; cependant, lorsque le contexte acoustique est très mauvais, les meilleures hypothèses peuvent se trouver exclues du faisceau de recherche. Ceci peut se produire plus fréquemment dans une configuration temps réel du système, pour laquelle les coupures sont plus strictes. Dans ce cas, une stratégie basée sur la "promotion" des hypothèses du faisceau coïncidant avec la transcription ne permet pas de récupérer ces erreurs. On peut quantifier de façon approximative la perte correspondant à cette situation en utilisant le moteur de reconnaissance pour faire un alignement forcé de la transcription exacte sur le signal.

Les expériences combinant l'interpolation des modèles de langage avec un alignement sur la transcription exacte sont présentées dans le tableau 3 .

TAB. 3: Résultats des expériences interpolant modèle de langage générique (ML-G) avec le modèle de langage appris sur la transcription exacte (ML-TrEx) et s'alignant sur la transcription exacte (alTrEx)

	Taux d'erreur
ML-G seul + alignement TrEx	6.1%
ML-TrEx seul + alTrEx	4.1%
ML-G70%+ML-TrEx30%+alTrEx	3.7%
ML-G50%+ML-TrEx50%+alTrEx	3.5%
ML-G30%+ML-TrEx70%+alTrEx	3.7%

Nous obtenons dans ce cas un taux d'erreur mots de 3.5%. Ce niveau d'erreur peut être considéré comme minimal pour une méthode ré-estimant les hypothèses concurrentes dans le faisceau d'hypothèses sans remettre en cause le contenu même de ce faisceau.

Le tableau 4 reprend les expériences précédentes mais en remplaçant la transcription exacte par la transcription approchée.

Le meilleur résultat est obtenu en combinant le modèle de

TAB. 4: Résultats des expériences avec interpolant le modèle de langage générique (ML-G) avec le modèle appris sur la transcription approchée (ML-TrEr), et s'alignant sur la transcription approchée (alTrEr)

	Taux d'erreur
ML-TrErr + alignement TrEr	9.9%
ML-G + alignement TrEr	7.7%
ML-G70%+ML-TrEr30%+alTrEr	7.2%
ML-G50%+ML-TrEr50%+alTrEr	7.4%
ML-G30%+ML-TrEr70%+alTrEr	8.6%

langage générique avec un poids de 70% avec le modèle appris sur la transcription approchée et en réalisant un alignement sur cette dernière. L'alignement fait descendre le taux d'erreurs mots jusqu'à 7.2%. Il permet d'apporter une information temporelle qui est mal prise en compte par le modèle de langage. L'utilisation d'un alignement DTW associé à l'interpolation des modèles montre à nouveau un gain. Cette expérience montre qu'un équilibre peut être atteint pour exploiter l'information approchée sans pour autant reproduire la majorité des erreurs qu'elle comporte. Les meilleurs résultats sont obtenus en utilisant le modèle de langage générique fortement pondéré par rapport au modèle de langage appris sur la transcription approchée et en alignant sur cette dernière. L'information erronée ne se trouve que dans la transcription sur laquelle le moteur essaye de s'aligner. Quand il ne trouve aucun alignement, il se replie exclusivement sur l'utilisation du modèle de langage générique. Par ailleurs, la légère interpolation avec le modèle de langage appris sur la transcription approchée permet de corriger certaines erreurs inhérentes au modèle de langage générique. Dans ces conditions, le système tire avantageusement parti de la transcription approchée lorsqu'elle est correcte et bascule en mode de reconnaissance automatique lorsque les observations acoustiques ne correspondent pas à la transcription proposée.

5. CONCLUSION

Nous avons proposé et évalué deux méthodes exploitant l'information contenue dans une transcription imparfaite pour améliorer les performances d'un SRAP. La première consiste à extraire du script l'information linguistique sous forme d'un modèle de langage trigramme appris sur la transcription approchée. Nos expérimentations montrent que l'interpolation de ce modèle avec le modèle de langage générique permet d'améliorer significativement le décodage. Il ne permet cependant pas de dépasser la qualité de la transcription approchée fournie, ce qui limite son intérêt. La seconde approche présentée consiste à orienter l'algorithme de recherche vers la transcription en synchronisant à la volée les hypothèses en cours d'évaluation et la transcription dont on dispose. Cette méthode permet de combiner efficacement les scores linguistiques avec les scores d'alignement. Partant d'un taux d'erreurs mots de 22.7%, notre méthode exploitant une transcription imparfaite permet de ramener ce taux jusqu'à 7.2%, montrant ainsi l'intérêt d'un alignement sur une transcription approchée : le décodage guidé permet de descendre le WER au dessous du WER de la transcription approchée (de 10.1% à 7.2%, soit 28% en relatif).

Un des intérêts du décodage basé sur A^* est la facilité avec laquelle des sources d'informations supplémentairespeuvent être intégrées au coeur même de l'algorithme de recherche. Ici, l'évaluation des hypothèses guidée par la transcription permet d'atteindre les objectifs fixés tout en accélérant le décodage. Ce gain au temps d'exécution est dû à la réduction de l'espace de recherche ainsi qu'à une meilleure anticipation des chemins optimaux, qui correspondent souvent aux hypothèses alignées. Cependant, le gain obtenu en terme de vitesse de décodage peut probablement être augmenté en introduisant plus tôt des heuristiques basées sur la transcription approchée, notamment au niveau de la sonde elle-même.

Bien que ces premiers résultats montrent l'intérêt d'un alignement sur une transcription approchée, ces expériences ont été effectuées dans des conditions contrôlées : niveau de bruit relativement réduit, transcription relativement proche de la transcription exacte, etc. Nous envisageons d'utiliser nos méthodes dans des conditions plus difficiles, par exemple sur le sous-titrage de films, ou l'alignement en temps réel de sous-titrages pour les pièces de théâtre.

RÉFÉRENCES

[1] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. 1994.

[2] Huang Chih-wei. Automatic closed caption alignment based on speech recognition transcripts. 2003.

[3] G. Linarès D. Massonié, P. Nocéra. Scalable language model look-ahead for lvcsr. *InterSpeech'05, Lisboa, Portugal*, 2005.

[4] Photina Jaeyung Jang and Alexander G.Hauptmann. Improving acoustic models with captioned multimedia speech. *IEEE International Conference on Multimedia Computing and Systems, Florence, Italy*, 1999.

[5] L. Lamel, J.L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic models training. *Computer Speech and Language*.

[6] G. Linarès, P. Nocéra, D. Matrouf, F. Béchet D. Massonié, and C. Fredouille. Le système de transcription du lia pour ester-2005. 2005.

[7] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A recursive algorithm for the forced alignment of very long audio segments. *International Conference on Spoken Language Processing*, 1998.

[8] Pascal Nocera, Georges Linares, and Dominique Massonié. Principes et performances du décodeur parole continue speeral. *XXIVées journées d'étude sur la parole*, 2002.

[9] Paul Placeway and John Lafferty. Cheating with imperfect transcripts. *Proceedings of ICSLP*, 1996.

Détection automatique d'opinions dans des corpus de messages oraux

Nathalie Camelin[1]*, Géraldine Damnati*[2]*, Frédéric Béchet*[1]*, Renato De Mori*[1] *

[1] LIA/CNRS - University of Avignon, BP1228 84911 Avignon cedex 09 France
[2] France Télécom R&D - TECH/SSTP/RVA 2 av. Pierre Marzin 22307 Lannion Cedex 07, France
{nathalie.camelin,frederic.bechet,renato.demori}@univ-avignon.fr
geraldine.damnati@francetelecom.com

ABSTRACT

Telephone surveys are often used by Customer Services to evaluate their clients' satisfaction and to improve their services. Large amounts of data are collected to observe the evolution of customers' opinions. Within this context, the automatization of the process of these databases becomes a crucial issue. This paper addresses the automatic analysis of audio messages where customers are asked to give their opinion over several dimensions about a Customer Service. Interpretation methods that integrate automatically and manually acquired knowledge are proposed. Experimental results, done on a database collected from a deployed Customer Service in real conditions with real customers are given.

1. INTRODUCTION

La détection d'opinions ou encore d'assertions objectives ou subjectives dans un texte est un domaine de recherche en pleine expansion [5, 1]. Du point de vue des utilisateurs, les deux principales applications de ce type de détection concerne d'une part l'analyse automatique d'opinions dans des messages contenant l'avis de consommateurs sur un produit ou un phénomène particulier [3] et d'autre part l'analyse de la subjectivité d'une phrase pour les systèmes de résumé automatique ou de question/réponse [4]. D'un point de vue scientifique, la problématique posée par la détection d'opinions se situe dans le cadre de la compréhension automatique de messages. Au niveau sémantique, ce problème constitue une possibilité d'aborder un niveau intermédiaire entre la simple détection des entités présentes et l'analyse sémantique complète du message, qui n'est pas envisageable sur des messages complexes.

La principale originalité de ce travail réside dans le type de message traité : nous abordons ici le problème de la détection d'opinions dans des messages oraux collectés auprès de *vrais* utilisateurs. Cette étude nous permettra ainsi, d'une part de tester la robustesse des processus de détection d'opinions aux erreurs de reconnaissance et aux disfluences ; et d'autre part de développer un module de transcription automatique de parole spécifique à ce type de corpus particulièrement difficile.

2. DESCRIPTION DU CORPUS

Les personnes sont invitées par un court message à appeler un numéro gratuit qui leur permet d'exprimer leur satisfaction vis à vis du service-client qu'ils ont récemment appelé. En composant ce numéro, le message vocal suivant les invite à laisser un message : *[. . .] Vous avez récemment contacté notre service clientèle. Nous souhaitons nous assurer que vous avez été satisfait de l'accueil et de la suite donnée à votre appel. Vous pourrez me laisser votre réponse après le top sonore. [. . .]*

Du fait que les messages ont été enregistrés à l'origine dans l'optique d'un traitement par opérateur, aucune consigne de nature à faciliter le traitement automatique n'a été donnée : pas de conseils sur le mode d'élocution, question ouverte et même incitation à laisser des commentaires. Ainsi, les messages recueillis sont *réalistes* et de longueur variable (d'une dizaine à plusieurs centaines de mots).Pour cette étude un ensemble de 1779 messages, collectés sur une période de 3 mois, a été transcrit manuellement au niveau mots, opinions et marqueurs (indication de disfluence et marqueurs discursifs).

Au niveau des opinions, quatre critères ont été retenus : l'*accueil*, l'*attente*, l'*efficacité* et un dernier critère regroupant le reste des critères évoqués par le locuteur : *autre*. Ces critères s'expriment chacun selon deux polarités différentes : *plus* ou *moins*. Cela fait un total de 8 étiquettes sémantiques ou *concepts*. Dans la transcription manuelle, au sein de chaque message, l'expression d'une opinion sur l'un de ces critères est indiqué par des balises. Nous disposons ainsi d'un corpus de 1079 segments, chacun porteur d'une ou plusieurs opinions particulières. Le but du traitement automatique est de retrouver ces segments et de les étiqueter avec la ou les opinions appropriées.

Nb concept par message	Répartition (% corpus)	Taille moyenne (nb mots)
0	19.2	61.0
1	51.3	40.3
2 et plus	29.5	60.8

TAB. 1: Répartition des messages dans le corpus en fonction du nombre de concepts exprimés

La taille moyenne des messages en fonction du nombre d'opinions exprimées est présentée dans le tableau 1. Même si les messages exprimant une seule opinion sont les plus courts, on voit qu'un message long n'est pas forcément le signe d'un nombre plus grand d'opinions exprimées, notamment les messages constitués uniquement de digressions et porteur d'aucune opinion sont en moyenne les plus longs. Un autre problème que pose ces messages est qu'un même concept peut être vu plusieurs fois dans un message avec des opinions contraires. Cela se ren-

*Travaux réalisés en collaboration avec France Télécom's R&D - contrat 021B178

contre quand la personne n'est pas entièrement satisfaite (*e.g.* :satisfaite du service-client mais pas du résultat) ou qu'une notion temporelle rentre en jeu dans son discours. Un exemple de message est donné dans le tableau 2.

oui c'est monsieur NOMS PRENOMS j'avais appelé donc le service client ouais j'ai été très bien accueilli des bons renseignements sauf que ça ne fonctionne toujours pas donc je sais pas si j'ai fait une mauvaise manipulation ou y a un problème enfin voilà sinon l'accueil était et les conseils très judicieux même si le résultat n'est pas n'est pas là merci au revoir

TAB. 2: Exemple de message contenant plusieurs opinions

3. MODÈLES DE LANGAGE SPÉCIFIQUES AUX OPINIONS ET SEGMENTATION AUTOMATIQUE

Du fait du degré de liberté laissé aux utilisateurs dans l'énoncé de leur message, on observe une assez grande dispersion dans la distribution des fréquences des mots. Ceci est d'autant plus le cas dans les portions des messages où les utilisateurs relatent l'origine de leur problème qui peut être de nature assez variée. Une fois les noms propres filtrés, le corpus d'apprentissage dans son ensemble contient 2981 mots différents pour un nombre total 51056 occurrences. Près de la moitié des mots n'apparaissent qu'une seule fois dans le corpus d'apprentissage, et la restriction du lexique aux mots d'occurrence supérieure ou égale à 2, conduit à un lexique de 1564 mots pour un taux de mots hors-vocabulaire égal à 2,8%. Un premier modèle de type bigram a donc été construit sur la base de ce lexique réduit. Aux mots du vocabulaire s'ajoutent des éléments spécifiques aux données, tels qu'un modèle de rejet particulier pour les noms propres ou encore une grammaire de numéros de téléphones. Ces éléments sont intégrés au modèle bigram.

Parallèlement, une première tentative de segmentation automatique a été réalisée, avec pour objectif de proposer un découpage des messages pour faciliter la tâche de classification en aval. L'idée est d'évaluer l'apport d'une segmentation a priori et non supervisée des messages en utilisant un automate bruit/parole pour détecter automatiquement les pauses réalisées par les locuteurs. Même s'il n'y a pas a priori de corrélation entre la présence de pauses et le changement de thématique, cette première approche a le mérite d'être simple à mettre en oeuvre et servira de baseline pour la suite de l'étude. Les segments isolés par l'automate bruit/parole sont soumis indépendamment les uns des autres au système de reconnaissance et les hypothèses de reconnaissance associées sont transmises aux modules de classification. Ce modèle portera le nom de *RECO1* dans la section 5.

Cette segmentation s'est avérée insuffisante. En effet, il subsiste d'une part des segments assez longs et porteurs de plusieurs expressions d'opinion (enchaînés sans pause). Il arrive d'autre part que des portions porteuses d'opinion soient tronquées par la segmentation automatique (si l'utilisateur hésite par exemple alors qu'il exprime une opinion). Du fait du nombre très important de disfluence au sein des messages, mais aussi souvent de la mauvaise qualité acoustique des messages, le taux d'erreur mot moyen obtenu avec ce modèle sur l'ensemble du corpus est de 58%. Ce taux très important est à relativiser car il inclus toutes les répétitions, hésitations et digressions effectuées par les utilisateurs.

Dans un deuxième temps, les problématiques de segmentation et de reconnaissance ont été intégrées à travers un nouveau type de modèle de langage. L'idée est de ne modéliser explicitement que les portions de messages porteuses d'opinion. Pour cela, un sous-corpus a été extrait pour chaque étiquette qui regroupe l'ensemble des segments associés à cette étiquette dans le corpus d'apprentissage initial. Un sous-modèle bigram a ainsi été estimé pour chaque étiquette à partir du sous-corpus associé. Par ailleurs, un modèle englobant de type bigram portant sur les étiquettes elles-mêmes a été estimé pour modéliser les enchaînements entre les différents segments d'opinion. Les portions qui ne correspondent à aucune expression d'opinion sont quant à elles modélisées par une boucle de phonèmes en contexte, sans contraintes a priori sur les enchaînements de phonèmes. L'ensemble est compilé au sein d'un unique modèle, appelé *RECO2* dans la présentation des expériences de la section 5 . La figure 1 présente ces trois types de modèles sur un exemple de message.

L'ensemble des segments extraits sur toutes les étiquettes représente environ 18700 occurrences de mots et le nombre de mots différents par sous-corpus ne dépasse pas 780 pour une moyenne de 470. Le premier intérêt est donc d'avoir réduit fortement le champ lexical. Par ailleurs, les messages se caractérisent globalement par un haut degré de disfluence. Or à nouveau, les parties les plus disfluentes ne sont pas celles où le locuteur exprime son opinion mais plutôt celles où il relatent l'origine de son problème initial. On observe ainsi une réduction du degré de disfluence dans les segments extraits. Ceci est illustré dans le tableau 3.

Indicateur	# messages	# segments
pauses remplies	6.1	5.0
faux départs	1.9	1.7
reprises	4.2	3.9
répétitions	2.0	2.3
marqueurs discursifs	4.3	1.2

TAB. 3: Pourcentage des indicateurs de disfluence dans le corpus global et dans le corpus extrait

Hormis les répétitions, qui ne sont pas les phénomènes les plus problématiques pour la reconnaissance, l'ensemble des indicateurs ont un pourcentage plus faible dans les segments d'opinion extraits. La baisse la plus significative concerne les marqueurs discursifs qui sont assez difficiles à modéliser du fait de la variété de leurs contextes d'apparition et qui peuvent perturber le traitement ultérieur des messages du fait de leur ambiguïté. Les mots "bon" ou "bien" par exemple peuvent à la fois être porteurs de sens pour une opinion et neutres quand ils sont employés pour articuler le discours.

4. CLASSIFICATION AUTOMATIQUE

La détection d'opinions dans un message peut se ramener à une tâche de classification : attribuer à un message une étiquette relative à l'expression d'une opinion particulière.

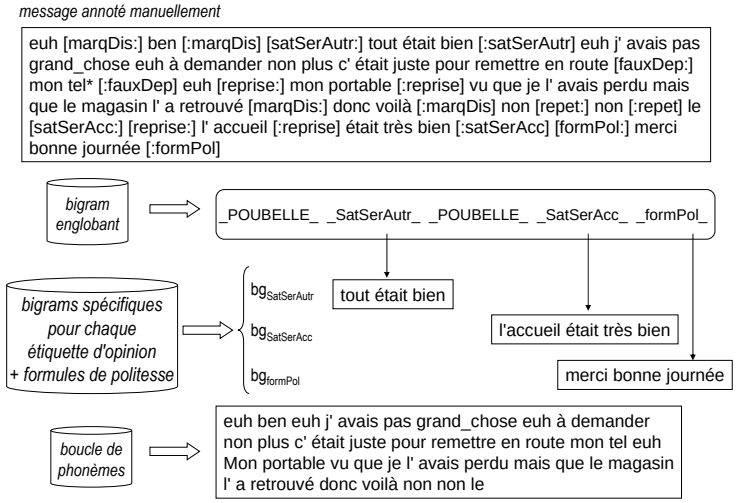


FIG. 1: Exemple de message annoté avec les 3 types de modèles de langage utilisés

Pour cette étape, nous utilisons une méthode de classification qui a prouvé son efficacité sur des tâches de classification de texte : les machines à support vectoriel ou *SVM*. L'implémentation des SVM utilisée dans cette étude est l'outil *SVMTorch* [2].

Comme mentionné précédemment, un message peut être caractérisé par plusieurs concepts (différents, identiques, contraires ...). Nous avons appris 8 classifieurs binaires, un par étiquette d'opinion, et l'étiquetage d'un message consiste à appliquer ces 8 classifieurs puis à en concaténer les décisions pour produire.

Les modèles sont appris sur le même corpus d'apprentissage que celui utilisé pour les modèles de langage présentés au paragraphe 3. *SVMTorch* est un classifieur qui prend en entrée un vecteur de données représentant le message à classer. Ce vecteur est de taille fixe pour un modèle considéré. Le choix des paramètres consiste donc à modéliser le message sous la forme d'un vecteur. De très nombreuses modélisations sont possible pour classer des données textuelles telles que la technique du *sac de mots*, la représentation par vecteurs de n-grammes, des n-grammes à trous, etc. Cette étude étant centrée sur l'impact de différents modèles de segmentation et de reconnaissance de parole sur les performances de classification, plus que sur l'étude de la modélisation pour les classifieurs, nous avons limité nos expériences aux deux représentation suivantes :

1. la modélisation la plus simple consiste à considérer comme vecteur d'entrée, le lexique complet de l'application. Un message est alors représenté par un vecteur de 2981 composantes. Les composantes non nulles de ce vecteur sont le nombre d'occurrences de chaque mot apparaissant dans le message. L'ordre des mots dans le message n'entre pas en jeu.

2. Une approche typique de l'analyse des opinions dans un texte consiste à créer un lexique contenant un ensemble de mots (appelés *seeds*) susceptibles d'exprimer une polarité positive ou négative [6]. Cette modélisation consiste ainsi à limiter le vecteur représentant un message aux seuls mots *seeds* du lexique.

Pour cette dernière modélisation, un ensemble de mots *polarisés* a été listé manuellement. Exemple : aberrant, compliments, discourtois, embêtement, ... Afin de généraliser la liste de mots polarisés obtenue, chaque mot est remplacé par son lemme. C'est cet ensemble de 565 lemmes polarisés que nous notons *seeds*. Le message est alors représenté par un vecteur de 565 composantes.

5. EXPÉRIENCES

Deux expériences sont faites en parallèle selon les deux types de paramètres choisis (mots ou *seeds*). Les résultats obtenus sont présentés par le calcul de la précision P, le rappel R et une combinaison précision/rappel : la F-mesure F.

Le corpus testé représente 33% des 1779 messages collectés et transcrits manuellement. Il est transcrit sous trois formes :
– *REF* : les messages sont transcrits manuellement ;
– *RECO1* : les messages sont transcrits automatiquement. Le modèle utilisé est de type bigram sur le lexique complet de l'application.
– *RECO2* : les messages sont transcrits automatiquement avec le modèle présenté en section 3

5.1. *Classification de messages sans segmentation préalable*

Les différents modèles appris sur le corpus d'apprentissage détectent les concepts recherchés sur le corpus de test. Les résultats sont indiqués dans le tableau 4.

Les résultats montrent bien la difficulté de la tâche. Ils sont d'ailleurs similaires à ceux obtenus dans ce type de caractérisation [1] avec une précision approchant les 60% sur du texte propre.

	mots			seeds		
(en%)	P	R	F	P	R	F
REF	59.2	37.3	45.8	57.9	43.6	49.8
RECO1	52.7	28.0	36.5	53.9	37.7	44.4
RECO2	51.5	34.9	41.6	52.3	40.0	45.4

TAB. 4: Résultat sur le message non-segmenté, référence manuelle (REF) et deux modèles de reconnaissance de parole (RECO1) et (RECO2)

Malgré un fort taux d'erreur mot, la détection des critères dans le message transcrit automatiquement ne fait perdre que 7 points de précision. Le rappel lui décroît de 10 points. L'utilisation des modèles *RECO2* au lieu des modèles *RECO1* permet de pallier cette perte de rappel avec moins de 3 points de différence avec le rappel obtenu sur le texte propre.

L'utilisation des seeds au lieu des mots permet d'améliorer la F-mesure sur le texte propre de 4 points. On remarque que c'est surtout le rappel qui tire avantage de ce type de paramètre. De même en ce qui concerne *RECO1* et *RECO2*, l'utilisation des seeds augmente le rappel de 10 points pour *RECO1* et de plus de 5 points pour *RECO2* et permet ainsi de passer d'une F-mesure de 36.52% à une F-mesure de 45.36% et de n'être plus qu'à 4 points de la F-mesure maximale obtenue jusqu'ici sur le texte propre.

5.2. Segmentation de messages

Les bons résultats obtenus avec *RECO2* et l'utilisation des seeds montrent que le traitement du message par segment permet de mieux cerner l'information pertinente. Comme pour *RECO2*, un sous-corpus ne contenant que les segments d'interventions porteurs de sens d'après l'annotation manuelle du corpus d'apprentissage, est extrait. Tous les modèles évoqués précédemment sont ré-appris sur ce sous-corpus. Le message d'entrée est lui aussi segmenté. Cette segmentation dépend du type de transcription du corpus.

– REF : segments obtenus manuellement, ce sont ceux porteurs de l'étiquette sémantique recherchée.
– RECO1 : segments obtenus automatiquement selon les pauses marquées par le locuteur durant son discours.
– RECO2 : segments obtenus automatiquement à chaque changement de modèles de transcription.

Chaque message est alors représenté par un ensemble de segment. Chaque segment est testé indépendamment par chaque modèle. L'ensemble des réponses obtenues pour un même message est ensuite rassemblé pour ne juger finalement que l'étiquette globale obtenue par le message sur l'ensemble de ses segments. Les résultats obtenus sont présentés dans le tableau 5.

	mots			seeds		
(en%)	P	R	F	P	R	F
REF	75.8	60.0	67.0	72.8	68.7	70.7
RECO1	49.0	33.9	40.0	48.1	43.9	45.9
RECO2	39.7	63.8	48.9	40.2	62.9	49.0

TAB. 5: Comparaison de 3 méthodes de segmentation de messages

Les résultats du corpus REF sont nettement améliorés avec une augmentation de la F-mesure de 20 points. Dans une moindre mesure, les résultats obtenus sur *RECO1* et *RECO2* sont eux aussi améliorés. On observe que c'est surtout le rappel qui est amélioré. En effet, l'apprentissage sur les segments porteurs de sens cible précisément les segments que l'on retrouve donc plus facilement. *RECO2* montre l'augmentation du rappel la plus remarquable allant de 20 à 30 points. Ceci s'explique par la construction même du message de *RECO2* qui recherche exactement les mêmes segments que ceux qui ont servis à l'apprentissage des modèles discriminants.

6. CONCLUSIONS ET PERSPECTIVES

La spécification du module de transcription à notre tâche ainsi que la recherche d'information pertinente pour la construction des modèles de classification et la représentation du message nous ont permis d'améliorer les premiers résultats.
En effet, les derniers résultats obtenus sur le texte propre sont acceptables dans ce type de tâche qui reste très difficile. Cette difficulté est amplifiée par la translation du problème de détection d'opinions sur des messages oraux énoncés par de *vrais* utilisateurs. Malgré le fort taux d'erreurs induit par cette parole spontanée, la détection d'opinions dans les messages issues du module de transcription obtient des résultats qui dépassent notre première baseline sur le texte propre.

Il s'agit maintenant de persévérer dans une recherche d'information plus pertinente en étendant par exemple les *seeds* à des patterns afin de capter le contexte et donc de mieux structurer l'information. L'intégration d'informations d'autres niveaux, par exemple prosodique, semble également indispensable, tant il est vrai que d'une part toute l'information portée par un message ne se résume pas à sa transcription lexicale, et que d'autre part la transcription automatique des messages est elle même peu fiable.

RÉFÉRENCES

[1] Yejin Choi, Claire Cardie, Ellen riloff, and Siddharth Patwardhann. Identifyind sources of opinions with conditional random fields and extraction patterns. In *Proceedings of HLT/EMNLP*, pages 355–362, Vancouver, 2005.

[2] Ronan Collobert, Samy Bengio, and Johnny Mariéthoz. Torch : a modular machine learning software library. In *Technical Report IDIAP-RR02-46, IDIAP*, 2002.

[3] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*, pages 339–346, Vancouver, 2005.

[4] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing*, 2003.

[5] Jayce Wiebe, Theresa Wilson, and Claire Cardie. Annotationg expressions of opinions and emotions in language. In *Language Resources and Evaluation*, volume 39, pages 165–210, 2005.

[6] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, pages 347–354, Vancouver, 2005.

Estimation rapide de modèles de Markov semi-continus discriminants

Georges Linarès, Christophe Lévy, Jean-christophe Plagniol

Laboratoire Informatique d'Avignon
339 Chemin des meinajaries, BP 1228, 84911 Avignon, France
{georges.linares, christophe.levy, jean-christophe.plaignol}@univ-avignon.fr

ABSTRACT

In this paper, we present a fast estimation rule for MMIE (Maximum Mutual Information Estimation) training of semi-continuous HMM (Hidden Markov Models).

The first experiments validate this method by comparing our fast MMI estimator (FMMIE) and the original one. We observe that, on a digit recognition task, FMMIE and full MMIE estimation obtain similar results, when our method decreases significantly the computational time.

Then, we incorporate our semi-continuous MMIE models in a real-time Large Vocabulary Continuous Speech Recognition (LVCSR) system. The evaluation corpora is extracted from the French BN Evaluation campaign Ester. The results show that the proposed MMIE models outperform significantly the system based on continuous models while remaining at the same level of complexity.

1. INTRODUCTION

Différentes voies ont été explorées dans le passé pour limiter les ressources utilisées par les modèles acoustiques, tant en terme d'espace mémoire que de temps de calcul. Des solutions basées sur une modélisation par modèles de Markov semi-continus (SCHMM, Semi Continuous Hidden Markov Models) ont étés proposées dans la littérature ([3], [9], [7]). Dans ces architectures, les modèles partagent un dictionnaire commun de gaussiennes, les états étant caractérisés par un vecteur de poids généralement estimé par maximisation de la vraisemblance. Cette mutualisation massive des paramètres permet de réduire de façon très significative l'espace mémoire requis par le stockage des modèles acoustiques et peut limiter les problèmes d'estimation liés à des tailles de corpus d'apprentissage limitées. Par contre, le gain obtenu en terme de temps de calcul est moins décisif, les méthodes de calcul rapide des vraisemblances sélection de gaussiennes permettant d'atteindre le temps réel sur des systèmes grand vocabulaire tout en préservant la précision des modèles à plusieurs millions de paramètres.

Plusieurs méthodes d'estimation du dictionnaire de gaussiennes d'un SCHMM ont été proposées : dictionnaires multiples, gaussiennes issues d'un jeu de HMM classique ([11]), *etc*. L'estimation des poids, quand à elle, est généralement effectuée par maximisation de la vraisemblance (MLE, Maximun Likelihood Estimation) alors que l'apprentissage discriminant par MMIE s'est généralisé pour l'estimation des HMM continus ([1]), malgré l'augmentation très sensible du temps de calcul requis par ce type de techniques.

Dans ce papier, nous présentons une méthode d'estimation rapide des poids d'un SCHMM par maximisation de l'information mutuelle. Nous reformulons les règles de réestimation proposées dans [8] dans le cadre spécifique des modèles semi-continus et nous proposons une heuristique qui permet une estimation très rapide des poids. Cette méthode est d'abord évaluée sur un système mot-isolé petit vocabulaire, puis sur un système de reconnaissance grand vocabulaire.

La première partie de cet article décrit les principes de l'estimation des paramètres par maximisation de l'information mutuelle. La seconde partie décrit des expériences confrontant l'approximation proposée à la règle initiale dans le cadre d'un système embarqué de reconnaissance de chiffres. Dans la troisième partie, la méthode proposée est évaluée en reconnaissance de parole continue grand vocabulaire, sur une tâche de transcription temps-réel d'émission radiophoniques.

2. MMIE RAPIDE POUR L'ESTIMATION DES SCHMM

L'estimation des modèles acoustiques par maximisation de l'information mutuelle a donné lieu à de nombreux travaux ces dernières années. Le principe général est de minimiser le risque d'erreur en maximisant l'écart de vraisemblance entre la bonne transcription et les mauvaises. La recherche des paramètres λ améliorant la capacité discriminante des modèles est réalisée par des algorithmes d'optimisation maximisant la fonction objective :

$$F_{mmie}(\lambda) = \sum_{r=1}^{R} log(\frac{P_\lambda(O_r|M_{w_r})P(w_r)}{\sum_{\tilde{w}} P_\lambda(O_r|M\tilde{w})} \qquad (1)$$

où w_r est la transcription correcte, M_w la séquence de modèles correpondant à l'hypothèse w, $P(w)$ la probabilité linguistique et O_r une séquence d'observations. Le dénominateur somme les produits des probabilités acoustiques et linguistiques sur toutes les hypothèses possibles.

Une des difficultés majeures rencontrées pour l'estimation des paramètres optimaux réside dans la complexité de la fonction objective qui intègre, dans son dénominateur, l'ensemble des chemins incorrects suceptibles d'être empruntés (et les séquences de modèles associées). Pour atteindre une complexité acceptable, il faut limiter le nombre de ces chemins, par exemple en ne gardant que

les n meilleurs issus d'un treillis de mots ou de phonèmes ([10]). L'estimation des modèles par MMIE reste néanmoins bien plus coûteuse que par MLE, qui ne nécessite pas l'évaluation d'hypothèses incorrectes.

Dans le cas particulier des modèles semi-continus, seule la ré-estimation des poids est nécessaire. Par ailleurs, le partage massif des gaussiennes permet de réduire considérablement la complexité de l'estimation des paramètres. En effet, le calcul des vraisemblances est limité au nombre réduit des gaussiennes du dictionnaire. D'autre part, la présence des mêmes composantes dans tous les états peut permettre une sélection directe des gaussiennes discriminantes. Nous mettons ce point en évidence en développant la formule de ré-estimation des poids proposée dans [8]. Dans cet article, les auteurs montrent que les poids $\tilde{c}_{jm}$ maximisant la fonction objective peuvent être obtenus par maximisation de l'expression suivante :

$$\sum_{j,m} \left[\gamma_{jm}^{num} log(\tilde{c}_{jm}) - \frac{\gamma_{jm}^{den}}{c_{jm}} \tilde{c}_{jm} \right] \quad (2)$$

où γ_{jm}^{num} et γ_{jm}^{den} sont respectivement les taux d'occupation estimés sur les exemples corrects (num) et incorrects (den) ; c_{jm}, le poids de la composante m de l'état j à l'itération précédente ; $\tilde{c}_{jm}$, le nouveau poids de la composante (j, m).

En optimisant chaque terme de cette somme pour un ensemble de poids fixé, la convergence peut être obtenue après quelques itérations. Chacun de ces termes étant convexe, la formule de mise à jour se déduit directement de l'équation précédente :

$$\tilde{c}_{jm} = \frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} c_{jm} \quad (3)$$

où γ_{jm}^{k} est la probabilité d'être dans la composante m de l'état j estimée sur l'ensemble de données Ω_k, qui regroupe les trames associées à l'état k. Ce taux d'occupation peut s'exprimer en fonction des vraisemblances $L()$:

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \frac{L(X|S_j)}{\sum_i L(X|S_i)} \frac{c_{jm} L(X|G_{jm})}{L(X|S_j)} \quad (4)$$

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{jm})}{\sum_i L(X|S_i)} \quad (5)$$

En isolant, dans le dénominateur, la vraisemblance de la trame X sachant l'état S_k, on obtient :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{jm})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \quad (6)$$

Dans des modèles semi-continus, les composantes gaussiennes G_{jm} sont indépendantes de l'état j ; dans ce cas, les taux d'occupations peuvent s'écrire :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{km})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \quad (7)$$

En notant $\epsilon_k = \sum_{i \neq k} L(X|S_i)$, le rapport des taux d'occupations devient :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X \in \Omega^j} \frac{L(X|G_{jm})}{L(X|S_j)+\epsilon_j}}{\sum_l \sum_{X \in \Omega^l} \frac{L(X|G_{lm})}{L(X|S_l)+\epsilon_l}} \quad (8)$$

Nous proposons d'approximer le rapport précédent par :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{c_{jm}}{\sum_l c_{lm}} \quad (9)$$

En introduisant cette approximation dans l'équation 3, on obtient la formule de ré-estimation rapide des poids :

$$\tilde{c}_{jm} = \frac{c_{jm}^2}{\sum_l c_{lm}} \quad (10)$$

Après ré-estimation, les vecteurs de poids de chaque état sont re-normalisés.

Cette fonction de mise à jour des poids à chaque itération ne nécessite pas l'estimation des vraisemblances sur le corpus d'apprentissage. En terme de temps de calcul, son coût est limité à celui de l'apprentissage du modèle MLE initial, les poids MMIE se déduisant directement des poids MLE (*cf.* éq. 10).

Dans la section suivante, nous comparons les résultats obtenus par cette méthode aux résultats obtenus par estimation MMIE standard.

3. VALIDATION EXPÉRIMENTALE EN PETIT VOCABULAIRE

De façon à valider expérimentalement la formule de ré-estimation proposée, nous avons entraîné des modèles semi-continus par MMIE avec les formules originelles, puis avec l'approximation proposée. Le système est évalué sur une tâche de reconnaissance de chiffres isolés, avec une faible quantité de données d'apprentissage, ce qui autorise un calcul exact des formules de ré-estimation.

La plateforme utilisée est celle développée dans [5], dans le cadre d'un système conçu pour la reconnaissance de petits vocabulaires sur un téléphone portable.

3.1. Système de base

Ce système a été développé pour la reconnaissance de chiffres sous la contrainte de ressources matérielles très limitées. Il repose sur une modélisation par SCHMM et un processus d'adaptation à l'environnement acoustique avec très peu de données. Un GMM (Gaussian Mixture Model) initial, dont seront dérivés les GMM d'état, a été appris avec le corpus BREF120 ([4]), sur une centaine d'heure de parole.

Le GMM initial est ensuite adapté par MAP (Maximum A Postériori) sur un corpus de taille réduite, mais néanmoins caractéristique des conditions de test. Ces ensembles d'adaptation et de test sont issus de la base BDSONS ([2]), qui a été divisée en 2 sous-ensembles afin d'obtenir un corpus d'adaptation (destiné à l'adaptation du GMM et à la ré-estimation des poids) et un corpus de test. Ces 2 ensembles contiennent respectivement 700 et 2300 occurences de chiffres. Les vecteurs acoustiques sont composés de 12 coefficents PLP, auxquels on a ajouté l'énergie du signal. Ici, les paramètres dynamiques (dérivées premières et secondes) ne sont pas utilisés.

3.2. Résultats

Le tableau 1 présente les performances du système dans une configuration de reconnaissance embarquée. A par-

tir du même dictionnaire de gaussiennes, 2 approches sont utilisés pour l'estimation des poids : MMIE avec les formules complètes de ré-estimation (*cf.* éq. 3) et la ré-estimation rapide que nous proposons (FMMIE).

TAB. 1: Taux d'erreur mot en fonction de la taille du GMM initial exprimée en nombre de gaussiennes (# GAUSS) pour les modèles MMIE et MMIE rapide (FMMIE). Mesure effectuée sur 2300 chiffres issus de la base BDSONS.

# GAUSS	MMIE	FMMIE
216	4,26%	4,09%
432	2,70%	2,39%
864	2,48%	2,57%
1728	2,30%	2,00%

Les résultats obtenus montrent que les deux algorithmes se comportent globalement de la même façon. Dans la plupart des cas, FMMIE obtient des résultats légèrement meilleurs mais toujours dans l'intervalle de confiance. Par ailleurs, alors que les performances du modèle MMIE standard s'améliorent régulièrement avec l'augmentation de la taille du GMM initial, FMMIE semble légèrement plus irrégulier. Là aussi, ces variations restent limitées et toujours incluses dans l'intervalle de confiance.

4. EVALUATION SUR UN SYSTÈME GRAND VOCABULAIRE

De façon à évaluer notre méthode d'estimation rapide des poids MMIE dans un contexte de LVCSR, nous avons estimé un modèle semi-continu compact qui a été intégré au système "Broadcast News" du LIA ([6]). Les expériences ont ensuite été menées en utilisant le cadre expérimental défini pour la tâche temps-réel de la campagne d'évaluation ESTER.

4.1. Estimation du HMM semi-continu

Construction du dictionnaire de gaussiennes Les gaussiennes du dictionnaire sont extraites d'un HMM continu de petite taille. Il s'agit de modèles indépendants du contexte avec 64 gaussiennes par état (C-CI). Les vecteurs acoustiques utilisés sont composés de 12 coefficients PLP (Perceptual Linear Prediction), de l'énergie du signal dans la fenêtre d'analyse, puis des dérivées premières et secondes de ces 13 premiers coefficients. Enfin, les paramètres sont centrés et réduits sur une fenêtre glissante de $5s$. La première étape du processus d'estimation consiste à entraîner ces modèles continus classiques, pour lesquels chaque GMM est estimé sur les données spécifiques de l'état par un algorithme de type EM (Expectation-Maximisation). Les gaussiennes issues de cet apprentissage sont ensuite regroupées dans le dictionnaire qui est constitué d'environ 7000 éléments. Le modèle initial a été entraîné sur le corpus d'apprentissage d'ESTER, qui contient 200 heures de parole transcrite, issue d'émissions radiophoniques du groupe Radio-France. L'essentiel du corpus est composé de journaux d'information, avec une diversité de locuteurs assez importante.

Estimation des modèles semi-continus par MLE L'étape suivante constiste à entraîner des modèles dépendant du contexte par ré-estimation du vecteur de poids de chaque état. Le jeu de modèles utilisé est constitué de 10000 triphones partageant un ensemble de 3600 états émetteurs. Le partage des états a été déterminé par un arbre de décision.

Pour chacun de ces 3600 états, nous estimons d'abord un vecteur de poids par MLE sur le corpus d'apprentissage aligné par un HMM continu modélisant le même jeu d'unités acoustiques. Nous obtenons alors un premier ensemble de modèles dépendants du contexte (SC0), partageant les 7000 gaussiennes du dictionnaire, et dont chaque état est caractérisé par un vecteur de 7000 poids.

Les composantes significatives des états de SC0 sont ensuites selectionnées de façon à ce que la somme de leur poids atteigne un seuil γ fixé *a priori*. Pour une valeur de γ à $0,9999$, le modèle contient 580000 poids, ce qui représente une moyenne de 161 gaussiennes indexées par état. On obtient un modèle (SC-MLE) de 1100000 paramètres, soit une complexité qui dépasse très largement celle du modèle continu dont est issu le dictionnaire de gaussiennes (C-CI).

Pour comparer cette approche avec les modèles continus, nous avons entraîné un autre modèle semi-continu par MLE. Cette fois çi, γ est choisi de façon à atteindre une complexité équivalente à celle du modèle continu C-CI. Pour une valeur de γ à $0,3$, nous obtenons un modèle compact (SC-MLE-C) contenant 610000 paramètres.

Estimation des modèles semi-continus par MMIE Avec des corpus de taille importante ou sur des tâches grand vocabulaire, l'estimation exhaustive du dénominateur de la fonction objective MMIE ne peut pas être réalisée dans un temps raisonnable. Nous utilisons directement l'estimateur rapide proposé qui sera comparé aux approches à base de HMM continus ou semi-continus estimés par MLE.

Les poids MMIE sont ré-estimés à partir du modèle SC0 décrit dans le paragraphe précédent. Les composantes les plus significatives sont ensuite sélectionnées, comme pour le modèle SC-MLE, en supprimant les gaussiennes de poids très faible. Les modèles obtenus sont composés d'environ 80000 poids, ce qui représente un moyenne de 22 gaussiennes par état.

On obtient finalement un modèle (SC-FMMIE) d'une complexité de l'ordre de 615000 paramètres pour une valeur de γ identique à celle utilisée pour le modèle SC-MLE (qui est, lui, composé de plus de 1 million de paramètres). De nombreuses composantes ne sont donc pas spécifiques à un état, ce qui est une conséquence naturelle de l'estimation par maximum de vraisemblance des GMM, mais aussi probablement lié au manque de précision des modèles initiaux (indépendants du contexte et composés de seulement 7000 gaussiennes).

4.2. Résultats

Le tableau 2 présente les résultats obtenus sur 1 heure de parole extraite du corpus de développement de la base ESTER.

On peut d'abord constater que le modèle semi-continu entraîné par MMIE (SC-FMMIE) améliore les performances du modèle continu indépendant du contexte (C-CI) de

3,1% (en absolu). Dans le même temps, le nombre de paramètres n'a augmenté que de 12,7%. Le modèle semi-continu compact entraîné par MLE (SC-MLE) obtient, lui, un taux d'erreurs nettement inférieurs au modèle continu (35,4% contre 43,3% de WER pour le C-CI).

La coupure appliquée au modèle SC-MLE ($\gamma = 0,3$) pour obtenir le modèle SC-MLE-C (dont la complexité est comparable aux modèles SC-FMMIE et C-CI) est très stricte. Elle conduit à une dégradation conséquente de la qualité de ces modèles ; le taux d'erreur atteint 50,8% (soit une augmentation absolu de WER de 7,5%). Sans cette coupure (modèle SC-MLE) le nombre de paramètre est bien plus élevé (+202% par rapport à C-CI), mais le taux d'erreur s'abaisse à 35,4% (-7.5% en absolu par rapport à C-CI). Néanmoins, ces performances restent éloignées de celles obtenues par un modèle continu de grande taille : en utilisant un modèle de plus grande complexité (1000 états émetteurs, 4,8 millions de paramètres), le taux d'erreur est de 29,1% (sur la même base de test).

TAB. 2: Résultats (en taux d'erreur-mot) et complexité (en nombre de paramètres, noté # PAR) des systèmes basés sur des HMM continus indépendants du contexte (C-CI), des SCHMM avec estimation MMIE (SC-FMMIE), MLE réduit à 610000 paramètres (SC-MLE-C) et MLE (SC-MLE). Evaluation portant sur 1 heure d'émission radiophonique issue du corpus ESTER.

	C-CI	SC-FMMIE	SC-MLE-C	SC-MLE
WER	43.3%	40.2%	50.8%	35.4%
# PAR	544k	615k	610k	1100k

5. CONCLUSION

Nous avons présenté une technique rapide de ré-estimation des poids MMIE dans le cadre de HMM semi-continus. Les évaluations comparatives ont montré que cet algorithme permet d'obtenir des résultats proches de ceux obtenus par la méthode initiale (et parfois légèrement meilleurs), malgré une relative instabilité. Cependant, son principal intérêt tient à la consommation très faible de temps CPU qu'elle requiert : partant de SCHMM estimés par MLE, le coût additionel d'estimation des poids MMIE est quasiment nul.

Les premières évaluations que nous avons menées en LVCSR sont encourageantes. Elles montrent le potentiel des SCHMM discriminants dans des contextes de ressources mémoires et CPU limitées. Nous envisageons de poursuivre cette étude dans deux directions. D'une part, des évaluations complémentaires doivent être menées de façon à mesurer les performances de cette technique en fonction de la complexité des modèles. D'autre part, la stratégie de constitution du dictionnaire de gaussiennes peut être affinée.

Enfin, ce type d'architecture pourrait permettre des adaptations rapides à l'environnement et/ou au locuteur, difficiles à intégrer dans des systèmes contraints par la vitesse de décodage ou par les ressources mémoire disponibles.

RÉFÉRENCES

[1] L.R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, pages 49–52, Tokyo, Japan, April 1986.

[2] R. Carré, R. Descout, M. Eskénazi, J. Mariani, and M. Rossi. The French language database : defining, planning and recording a large database. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1984)*, pages 324–327, San Diego, California, USA, March 1984.

[3] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, and Ronald Rosenfeld. The SPHINX-II speech recognition system : an overview. *Computer Speech and Language*, 7(2) :137–148, 1993.

[4] L.F. Lamel, J.L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'1991)*, pages 505–508, Gênes, Italie, September 1991.

[5] C. Lévy, G. Linarès, P. Nocera, and J.F. Bonastre. *Embedded mobile phone digit-recognition*, chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2. Springer Science, H. Abut, J.H.L. Hansen and K. Takeda edition, 2006, *à paraître*.

[6] G. Linarès, P. Nocéra, D. Matrouf, F. Béchet D. Massonié, and C. Fredouille. Le système de transcription du lia pour ester-2005, 2005.

[7] Brian Kan-Wing Mak. Towards A compact speech recognizer : Subspace distribution clusteringhidden markov model. Technical Report CSE-TH-98-001, 20, 1998.

[8] D. Povey and P. Woodland. Frame discrimination training of hmms for large vocabulary speech recognition, 1999.

[9] T. Vaich and A. Cohen. Comparison of continuous-density and semi-continuous hmm in isolated words recognition systems. In *EUROSPEECH'99*, pages 1515–1518, 1999.

[10] V. Valtchev, J. Odell, P. Woodland, and S. Young. Mmie training of large vocabulary recognition systems, 1997.

[11] K.F. Lee X. D. Huang and H. Hon. On semi-continuous hidden markov modeling. In *ICASSP'90*, pages 689–692, 1990.

Deux stratégies articulatoires pour la réalisation du contraste acoustique des sibilantes /s/ et /ʃ/ en français

Martine Toda

Laboratoire de Phonétique et Phonologie, Université Paris III – CNRS UMR 7018
19, rue des Bernardins, 75005 Paris, France
martine_toda@yahoo.fr
http://www.cavi.univ-paris3.fr/ilpga/ed/student/stmt/

ABSTRACT

This paper reports two articulatory strategies used in the realization of /s/ - /ʃ/ contrast in French from the observation of the MRI data of seven native speakers. These strategies are: 'tongue position adjustment' and 'tongue shape adjustment'. By examining the articulation of the subjects whose frication noise was 'deviant', it appeared that they already used all the possibilities for compensation within their articulatory strategy. A better normalization of their frication noise would have required complex gestures (e.g. tongue backing *and* doming), which are presumably avoided by virtue of articulatory economy.

1. INTRODUCTION

Les fricatives sibilantes sourdes /s et ʃ/ en français se différencient du point de vue articulatoire par leur lieu d'articulation, qui peut être décrit, respectivement, par les traits distinctifs [coronal ; + antérieur] et [coronal ; - antérieur]. Cependant, une grande part de variabilité inter-individuelle est observée quant à la réalisation dentale ou alvéolaire et apicale ou laminale des consonnes coronales (Dart, [1]) Une importante variation inter-individuelle est également observée dans la fréquence du bruit de friction, (Tabain, [2]), ce qui pourrait résulter, en partie, de la différence d'articulation, mais aussi de la différence de morphologie du palais (Toda, [3]).

On peut alors se demander pourquoi une telle variation acoustique et articulatoire peut exister, puisqu'elle est propre à compromettre l'intelligibilité de la parole.

Une première hypothèse serait que, dû à une limitation articulatoire quelconque, les locuteurs « déviants » ne peuvent se rapprocher davantage de la cible acoustique idéale. Cette variation individuelle, toutefois, ne constituerait pas un handicap à l'intercompréhension : en effet, le bruit de friction n'est pas le seul indice acoustique employé pour identifier les sibilantes (Mann & Repp [4]). De plus, les auditeurs exercent une normalisation en fonction du locuteur (*ibid.*). Dans ce cas de figure, il suffirait donc que le contraste acoustique soit suffisant à l'intérieur des productions de chaque locuteur pour que les sibilantes soient correctement identifiées.

Une autre hypothèse serait que le contraste entre les deux sibilantes est suffisamment grand, et que la variation inter-locuteurs n'occupe qu'une proportion mineure sur le plan perceptif. Cette hypothèse est motivée par l'aspect discontinu du contraste entre /s/ et /ʃ/ (Perkell *et al.*, [5]). En effet, dans un continuum articulatoire allant de /s/ à /ʃ/, le recul du point de constriction entraîne un changement brusque du volume de la cavité en avant de la constriction, par la création d'une cavité sublinguale au moment où la langue cesse d'être en contact avec les incisives inférieures. Si la fréquence de résonance des fricatives sibilantes est conditionnée par la taille de la cavité antérieure, la différence acoustique entre les deux sibilantes qui en résulte serait également de nature discontinue. Il est possible de penser que les auditeurs acquièrent une sensibilité catégorielle en conséquence. Ainsi, la variation individuelle serait ignorée au profit de la différence qui oppose /s/ et /ʃ/.

L'objectif de la présente étude est, premièrement, d'examiner quels sont les gestes articulatoires mis en œuvre dans la réalisation du contraste entre /s/ et /ʃ/. En second lieu, en mettant en rapport les données acoustiques et articulatoires de sept locuteurs, nous tenterons de déterminer d'où vient la variation acoustique et dans quelle mesure cette variation est simplement négligée, ou inévitable.

2. MÉTHODE

Les données articulatoires et acoustiques ont été obtenues à ATR, Kyoto, Japon. Sept chercheurs et stagiaires français (six hommes et une femme), locuteurs natifs de français et âgés de 21 à 24 ans ont participé à l'expérience.

2.1. Données articulatoires

La morphologie du conduit vocal a été obtenue grâce à l'imagerie par résonance magnétique (IRM). Les sujets étaient invités à produire les fricatives cibles de manière soutenue sans reprise de souffle pendant les séquences d'acquisition qui duraient environ 23 secondes. Les fricatives sibilantes /s/ et /ʃ/ étaient présentées à l'intérieur de mots. Les sujets étaient invités à produire ces sons « exactement comme lorsqu'ils sont contenus dans ces mots ». Trois contextes différents étaient proposés pour /s/

et /ʃ/, ce qui permettait de contrôler la pertinence des productions. Seules les données pour les contextes « Assam » et « achat » seront analysées. De plus, une séquence d'acquisition spéciale a permis d'obtenir la forme des dents par contraste avec les tissus mous environnants. Par la suite, les données des dents ont été insérées dans les données des fricatives à l'aide d'un logiciel de traitement de données volumétriques.

La description articulatoire des sibilantes est basée sur l'observation du plan médio sagittal. De plus, l'aire du conduit vocal sur le plan coronal depuis les lèvres jusqu'à la région vélaire a été mesurée tous les millimètres pour donner lieu à une fonction d'aire.

2.2. Données acoustiques

La séance d'enregistrement avait lieu avant l'acquisition des données IRM et permettait au sujet de s'entraîner, de même qu'à l'expérimentateur de contrôler les productions soutenues en vue de la séance IRM. Dans la présente étude, nous considérerons les fricatives sibilantes contenues dans les mots « Assam » et « achat » produits à l'intérieur d'un énoncé. La portion centrale des fricatives, visiblement stable sur le spectrogramme, a été utilisée pour calculer le spectre moyen du bruit de friction (10 fenêtres de 8 millisecondes s'échelonnant sur une portion de 62 millisecondes). Chaque spectre est caractérisé par sa fréquence de coupure, estimée visuellement sur le spectre.

3. RÉSULTATS

3.1. Deux stratégies articulatoires

L'observation des profils sagittaux, ainsi que des fonctions d'aire, suggère l'existence de deux stratégies articulatoires. Quatre locuteurs présentent une stratégie de « recul de la langue » entre /s/ et /ʃ/ (Figure 1 ; D, Y, S2 et J). La forme de leur langue est quasi inchangée entre les deux fricatives et la partie qui forme la constriction tend à être la même pour /s/ et /ʃ/ (predorsum, lamina ou apex selon le locuteur). Trois de ces locuteurs (D, S2 et J) présentent clairement une protrusion des lèvres dans /ʃ/ par rapport à /s/.

Les trois autres locuteurs (Figure 1 ; S1, V1 et V2) présentent une stratégie de « déformation de la langue » entre /s/ et /ʃ/. La partie de la langue qui forme la constriction n'est pas la même entre /s/ et /ʃ/ (apex vs. lamina ; lamina vs. predorsum, etc.). Sur la fonction d'aire (Figure 2 ; S1 et V1), on peut remarquer que /ʃ/ est palatalisé (le conduit vocal reste étroit en arrière de la constriction) en comparaison à /s/, et à la différence des locuteurs de la première stratégie (Figure 2 ; S2 et J). Contrairement à la description habituelle du /ʃ/ français, aucun des locuteurs de cette deuxième stratégie ne présente de protrusion labiale.

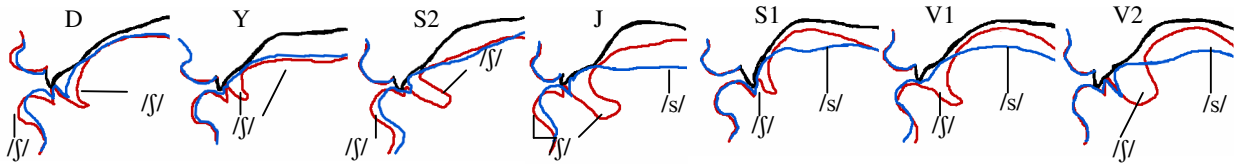


Figure 1 : contours sagittaux superposés (bleu : /s/ ; rouge : /ʃ/). Les quatre locuteurs de gauche présentent une stratégie « recul de la langue », tandis que les trois locuteurs de droite présentent une stratégie « déformation de la langue ».

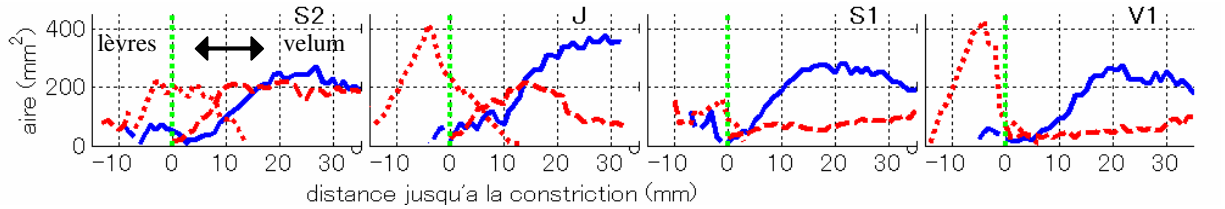


Figure 2 : fonction d'aire de quatre locuteurs (ligne pleine : /s/ ; pointillée : /ʃ/). Les fonctions d'aire sont alignées à la constriction maximale. La cavité antérieure de /ʃ/, comprenant une cavité sublinguale, est représentée en pointillés fins.

3.2. Aspects acoustiques

Aucun indice acoustique ne varie systématiquement avec les stratégies « recul » ou « déformation ».

Transitions formantiques

Les transitions formantiques (Figure 3) F1 et F2 présentent globalement les mêmes caractéristiques entre tous les locuteurs. F3 présente des schémas différents selon la consonne et le locuteur.

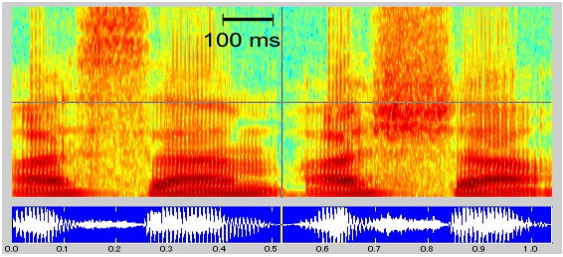


Figure 3 : spectrogramme (jusqu'à 8000 Hz) des mots « assam » et « achat », locuteur V1.

Pour /s/, F1 est descendant vers la fricative, F2 montant avec une cible allant de 1400 à 1800 Hz selon le locuteur, et F3 constant autour de 2700 Hz, sauf pour le locuteur D où il est montant pour atteindre 2600 Hz. Pour /ʃ/, F1 est descendant, F2 montant avec une cible allant de 1600 à 2000 Hz, et F3 constant pour les locuteurs S2 et V2 (aux alentours de 2500 Hz), montant pour les locuteurs J, S1, V1 et Y (cible vers 2900 Hz), et enfin, descendant pour le locuteur D (cible vers 2300 Hz).

Bruit de friction

La fréquence de coupure du bruit de friction (Figure 4) est, quant à elle, très variable selon le locuteur (Figure 5).

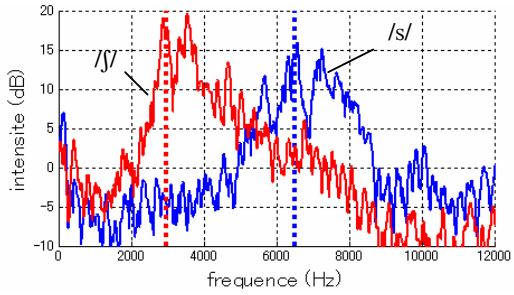


Figure 4 : spectre moyen et fréquence de coupure du bruit de friction pour /s/ (bleu) et /ʃ/ (rouge), loc. V1.

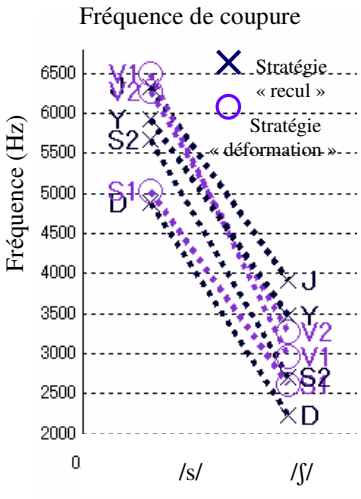


Figure 5 : Fréquence de coupure du spectre de bruit. Les croix indiquent les locuteurs de la stratégie « recul de la langue » ; les cercles, les locuteurs de la stratégie « déformation de langue ».

Le formant de la fréquence de coupure correspond au F3 (D, S2 et S1) ou au F4 (J, Y, V1 et V2) de la voyelle /a/ pour /ʃ/, et à des formants égal ou supérieur à F5 pour /s/. En ce qui concerne /ʃ/, cela suggère que la cavité responsable de la résonance principale du bruit de friction n'est pas la même selon les locuteurs. Pour certains, elle pourrait provenir du canal palatal, et pour d'autres, de la cavité antérieure comprenant la cavité sublinguale. De plus, on peut noter qu'il n'y a pas de correspondance systématique entre le formant excité par la source de bruit et la direction de la transition F3. Les valeurs les plus « déviantes » de la fréquence de coupure concerne le /s/ des locuteurs D (- 1,44 σ) et S1 (- 1,20 σ) et le /ʃ/ des locuteurs J (+ 1,55 σ) et D (- 1,41 σ). La différence entre le /s/ le plus « grave » (D) et le /ʃ/ le plus « aigu » (J) est d'à peine 1000 Hz, ce qui est peu par rapport à la différence à l'intérieur de chaque catégorie, qui avoisine les 1600 – 1700 Hz. Néanmoins, il existe à l'intérieur de tous les locuteurs, y compris « déviants », une distance importante entre /s/ et /ʃ/, qui est de l'ordre de 2500 Hz.

4. DISCUSSION

Le mouvement du F3 observé sur le spectrogramme pourrait avoir comme origine deux mécanismes distincts. Premièrement, en supposant, d'après Fant [6], que le F3 de la voyelle [a] correspond à une résonance de la cavité postérieure, le mouvement du F3 pourrait indiquer le changement de longueur de cette cavité. Cependant, un deuxième mécanisme pourrait aussi être responsable de ce mouvement. En effet, pour /ʃ/, la cavité sublinguale pourrait être considérée comme une cavité branchante. Une telle cavité introduit un pôle et un zéro libres, susceptibles de perturber la structure formantique des voyelles adjacentes. Sa taille peut raisonnablement donner lieu à cette structure dans la région du F3 de la voyelle, pour occasionner un changement d'affiliation de cavité du F3 et un mouvement apparent. Par conséquent, il est difficile d'interpréter le mouvement du F3, surtout en l'absence de données dynamiques.

Quant au bruit de friction, pourquoi les locuteurs ne pourraient-ils pas « faire mieux » pour estomper cette variation ? Celle-ci est d'autant plus intrigante que les mécanismes de compensation mis en œuvre pour les voyelles sont bien connus. Si les locuteurs n'effectuent pas de compensation, serait-ce parce qu'ils n'en ont pas les moyens ?

En ce qui concerne le locuteur S1, on observe que la cavité orale antérieure du /s/ est très réduite (voire inexistante). Nous supposons qu'une plus forte déformation de la langue n'aurait pas d'effet sur la taille de cette cavité. Pour augmenter la fréquence de résonance de la cavité antérieure, il lui faudrait raccourcir sa cavité labiale, ce qui n'est pas observé ici. En effet, un tel geste reviendrait à ajouter une dimension articulatoire supplémentaire dans la palette des traits disponibles (cf. Clements, [7]).

Quant au /ʃ/ du locuteur J, on observe une articulation laminale post-alvéolaire avec protrusion labiale. Pour abaisser davantage la résonance de la cavité antérieure, le geste labial étant déjà employé, la seule possibilité serait de reculer encore le lieu d'articulation. Un tel geste ne semble pas impossible à réaliser étant donné l'articulation des locuteurs V1 et V2. Cependant, pour parvenir à cette configuration, ces locuteurs bombent le

dos de leur langue. À l'intérieur de la stratégie « recul de la langue », le locuteur J aurait donc épuisé toutes ses cartes.

Le locuteur D présente également une stratégie de « recul ». Pour diminuer la taille de la cavité antérieure dans /s/, une possibilité serait d'avancer la langue. Or, elle semblerait déjà avoir atteint la limite antérieure étant donnée sa forme bombée. Une déformation de la langue serait donc requise pour l'avancer davantage.

L'examen de ces cas particuliers nous pousse à penser que les locuteurs tendent à rejeter les articulations complexes. D'après ce comportement, nous pouvons supposer que la différence minimale entre les /s/ les plus graves et les /ʃ/ les plus aigus est malgré tout suffisamment grande de sorte qu'il n'y a pas d'impératif à normaliser les propriétés du bruit de friction en fournissant des efforts supplémentaires.

Enfin, nous n'avons pas pu déterminer pourquoi les locuteurs adoptent l'une ou l'autre des stratégies articulatoires. On peut songer à trois principales causes : la recherche d'une structure résonantielle contribuant au contraste phonémique ; les contraintes liées à la création du bruit de friction strident ; et la coarticulation. Pour éclaircir l'influence du contexte phonétique, des études dynamiques sont indispensables. En ce qui concerne la source, on sait que la distance et l'angle de la constriction par rapport aux dents ou à la paroi qui sert d'obstacle sont primordiaux à la création du bruit strident qui caractérise les sibilantes (Pastel [8]). Cette source se situerait près des dents pour /s/ (Narayanan & Alwan, [9]; Shadle; [10]) et également près des dents (Shadle, [10]) pour /ʃ/ avec éventuellement une deuxième source autour de la constriction [9]. Toutefois, ces études ne comportent qu'un nombre restreint de modèles. Selon la morphologie buccale des locuteurs, des gestes différents pourraient être requis pour remplir les conditions aérodynamiques nécessaires. Il a par ailleurs été montré qu'un changement mineur dans la morphologie de la cavité antérieure pouvait induire une différence importante de la source du bruit (Shadle, 10). Le choix de la stratégie « recul » ou « déformation » pourrait donc servir, de manière non négligeable, à manipuler la source de bruit, au même titre que la structure résonantielle.

5. CONCLUSION

Cette étude avait comme objectif de mettre en évidence la façon dont est réalisé le contraste entre /s/ et /ʃ/ en français, et d'y chercher les causes de la variation acoustique. Deux stratégies articulatoires permettant de réaliser ce contraste ont été observées d'après les données IRM de sept locuteurs. Ces stratégies font appel à un seul geste de la langue à la fois (recul ou déformation), accompagné ou non d'un geste labial (ne comprenant qu'une seule modalité : la protrusion). Les locuteurs dont le bruit de friction est déviant ne pourraient pas « faire mieux », car ils semblent avoir épuisé toutes les possibilités offertes par leur stratégie articulatoire. En effet, ils évitent de faire appel à une dimension articulatoire supplémentaire. Pourtant, on doute que les locuteurs en soient physiquement incapables, car il existe des langues dans le monde avec un inventaire plus riche de sibilantes, dont les traits distinctifs comprennent à la fois le lieu d'articulation et la forme de la langue ou des lèvres. Il serait intéressant d'étudier comment les locuteurs réalisent ces contrastes dans ces langues (ex. mandarin, polonais, russe).

Remerciements Nous remercions vivement tous nos locuteurs, ainsi que l'équipe du Brain Activity Image Centre d'ATR pour l'acquisition des données IRM.

BIBLIOGRAPHIE

[1] S. N. Dart. Comparing French and English coronal consonant articulation. In *Journal of Phonetics*, volume 26, pages 71-94, 1998.

[2] M. Tabain. Variability in fricative production and spectra : implications for the hyper- and hypo- and quantal theories of speech production. In *Language and Speech*, Volume 44 (1), pages 57-94, 1998.

[3] M. Toda. Effect of palate shape on spectral characteristics of coronal fricatives. Communication orale, « Conference on turbulences », 13-14 octobre 2005, ZAS, Berlin. http ://www.zas.gwz-berlin.de.

[4] V. A. Mann & B. H. Repp, Influence of vocalic context on perception of the [s]-[ʃ] distinction. In *Perception & Psychophysics*, volume 28 (3), pages 213-228, 1980.

[5] J. S. Perkell, S. E. Boyce & K. N. Stevens. Articulatory and acoustic correlates of the [s - š] distinction. In *Speech Communication Papers*, 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), pages 109-113, 1979.

[6] G. Fant. Formants and cavities. In *Proc. 5th ICPhS*, Münster, pages 120-141, 1964. (Publié par S. Karger, Basel/New York, 1965.)

[7] G. N. Clements. Testing feature economy. In *Proc. 15th ICPhS*, Barcelona, pages 2785-2788, 2003.

[8] L. M. P. Pastel. *Turbulent noise sources in vocal tract models*. M. S. dissertation, Massachusetts Institute of Technology, 1987.

[9] S. Narayanan & A. Alwan. Noise source models for fricative consonants. In *IEEE transactions on speech and audio processing*, volume 8, numéro 2, pages 328-344, 2000.

[10] C. Shadle. The effect of geometry on source mechanisms of fricative consonants. *Joural of Phonetics*, volume19, pages 409-424, 1991.

Étude acoustique et articulatoire de la parole Lombard : Effets globaux sur l'énoncé entier

Maëva Garnier[1], Lucie Bailly[1,2], Marion Dohen[2], Pauline Welby[2], Hélène Lœvenbruck[2]

[1]Laboratoire d'acoustique musicale, CNRS UMR 7604, UMPC, Ministère de la culture, 11 rue de Lourmel, 75015 Paris
[2]Institut de la communication parlée, CNRS UMR 5009, INPG, Univ. Stendhal, 46 av. F. Viallet, 38031 Grenoble
Mél : garnier@lam.jussieu.fr, {lucie.bailly, pauline.welby, helene.loevenbruck}@icp.inpg.fr, marion.dohen@gmail.com

ABSTRACT

This study aims at characterizing the articulatory modifications that occur in speech in noisy environments, and at examining them as compensatory strategies. Audio, EGG and video signals were recorded for a female native speaker of French. The corpus consisted of 33 short sentences with a Subject-Verb-Object (SVO) structure. The sentences were recorded in three conditions : silence, 85dB white noise, 85dB "cocktail party" noise. Labial parameters were extracted from the video data. The analyses enabled us to examine the effect of the type of noise and to show that hyper-articulation concerns lip aperture and spreading rather than lip pinching. The analysis of the relationship between acoustic and articulatory parameters show that this speaker especially adapts to noise not only by talking louder or increasing vowel recognition cues but also by increasing spectral emergence.

1. INTRODUCTION

Les environnements bruyants affectent les situations de communication car ils perturbent à la fois la perception que le locuteur a de sa propre voix et son intelligibilité vis-à-vis de son interlocuteur. Pour compenser ces perturbations, le locuteur modifie sa façon de parler. Ce phénomène est connu sous le nom d'effet Lombard et a été principalement décrit des points de vue acoustique et phonétique [1-4]. Dans cette étude, nous nous intéressons aux modifications articulatoires concomitantes de l'effet Lombard. Nous cherchons à caractériser ces modifications articulatoires par rapport à la parole produite dans le silence. Existe-t-il une hyper-articulation? Est-elle conjointe à la modification de certains paramètres acoustiques ? Quels sont les paramètres articulatoires concernés par ces modifications ? Enfin, l'articulation de la parole Lombard dépend-elle du type de bruit ? À notre connaissance, quelques études se sont intéressées à l'articulation de certains types de parole "hyper" [5-7] mais très peu d'études concernent spécifiquement l'articulation de la parole Lombard [8-9].

2. MATÉRIEL ET MÉTHODE

2.1. Corpus

Les 33 phrases du corpus ont été construites sur une même structure Sujet-Verbe-Objet et selon un enchaînement de syllabes de structure CV, afin de simplifier la segmentation acoustique [10] (cf. (1) pour exemple). Seules des consonnes sonores ont été choisies afin de minimiser les perturbations de la courbe de F0.

(1) Le monument et les moulins rallient la vallée.

2.2. Enregistrement audiovisuel

Nous avons enregistré simultanément les signaux audio, électroglottographiques (EGG) et articulatoires d'une locutrice française, n'ayant aucune connaissance du protocole ou de l'effet Lombard . Elle avait pour consigne de lire les phrases en s'adressant à l'expérimentatrice située à 2m en face d'elle. Les données articulatoires ont été extraites d'enregistrements vidéo (25 images/s) de face des lèvres de la locutrice, en utilisant un dispositif labiométrique développé à l'ICP [11]. Dans le cadre de cette étude, nous nous sommes concentrées sur l'analyse de **l'étirement (A)** et de **l'ouverture (B)** des lèvres, ainsi que de **l'aire intéro-labiale (S)** (cf. Figure 1). Pour ces trois paramètres articulatoires, nous avons examiné à la fois les maxima d'amplitude des mouvements (**max**) ainsi que leur variation plus globale (**glob**) bien représentée par l'aire sous la courbe d'évolution de A, B, ou S en fonction du temps [12, p113-114].

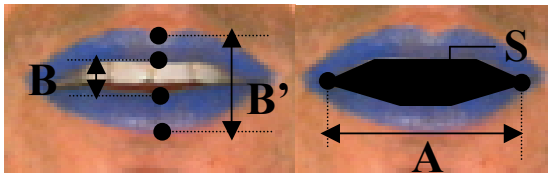


Figure 1 : Représentation des paramètres articulatoires

Nous avons également examiné un quatrième paramètre : le pincement, défini comme la compression des lèvres au moment de leur fermeture sur les segments [m], et donc équivalent au paramètre B' lorsque B est égal à 0 (cf. Figure 1). On observe deux types de pincement : **le pincement protrus** pour lequel la surface labiale visible augmente au moment de la compression, et **le pincement avalé** correspondant à un écrasement des lèvres vers l'intérieur de la bouche avec rétrécissement de la surface labiale visible.

Le signal audio a été enregistré à l'aide d'un microphone AKG placé à 20cm des lèvres, et le signal EGG à l'aide d'un électroglottographe EG2. Ils ont été numérisés à 44.1kHz sur 16bits. Deux environnements bruyants : un bruit blanc **(bb)** et un bruit de "cocktail party" **(cktl)**, tirés de la base BD_Bruit [13], ont été diffusés à 85dB (intensité mesurée au niveau des oreilles de la locutrice) par deux haut-parleurs placés à 2m de la locutrice et espacés de 2m. Leur spectre moyenné est représenté Figure 2. La locutrice a d'abord été enregistrée en situation silencieuse de référence, puis dans chacun des deux environnements bruyants.

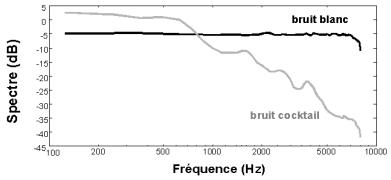


Figure 2: Spectre moyenné des deux types de bruit étudiés.

Les données audio ont été traitées pour supprimer le bruit (algorithme de [14]) et étiquetées sous Praat en déterminant les frontières des phrases, des groupes nominaux du sujet et de l'objet des phrases, et de leurs segments. La fréquence fondamentale **(F0)** et le quotient ouvert **(Oq)** ont été mesurés à partir du signal EGG par une méthode d'autocorrélation.

3. RÉSULTATS

Nous avons adopté dans la suite de l'article la notation suivante pour exprimer les résultats statistiques : * pour p<0,05, ** pour p<0,01, *** pour p<0,001, et ns pour p>0,05 (non significatif).

3.1. Articulation des énoncés dans le bruit

Amplitude des mouvements articulatoires

On observe une augmentation significative de A, B et S dans le bruit, pour les deux types de bruit, et une diminution significative du pincement protrus dans le bruit blanc (cf. Table 1). On remarque un effet significatif du type de bruit sur l'augmentation de tous les paramètres articulatoires, excepté le pincement avalé. Cette augmentation de l'amplitude des paramètres articulatoires que nous nommerons par la suite "hyper-articulation" est globalement plus marquée dans le bruit cocktail que dans le bruit blanc.

Table 1 : Évolution de l'amplitude des mouvements articulatoires dans les deux environnements bruyants.

		Δ dans bb	Δ dans cktl	Δ cktl-Δbb
A	max	+3,4% ***	+6,5% ***	+3% **
	glob	+8,9% ***	+14,2% ***	+5,3% ***
B	max	+19,5% ***	+27,5% ***	+7,9% ***
	glob	+16,0% ***	+27,1% ***	+11,2% ***
S	max	+28,4% ***	+42,3% ***	+13,9% ***
	glob	+25,3% ***	+ 43,6% ***	+18,3% ***
Pincement protrus	max	-16,3% ***	-11,0% *	+27,3% ***
	glob	-36,7% ***	-10,2% *	+26,5% ***
Pincement avalé	max	+83,0% **	+95,5% **	+12,5% -
	glob	+149,2% **	+141,1% **	-8,1% ns

Pics de vitesse des mouvements articulatoires

Les pics de vitesse (Vmax) augmentent significativement dans le bruit pour A, B et S, quel que soit le type de bruit, et diminue significativement pour le pincement protrus dans le bruit blanc (cf. Table 2). On observe une différence significative entre les deux types de bruit au niveau de l'augmentation des pics de vitesse de B, S et du pincement protrus. Les pics de vitesse augmentent davantage dans le bruit cocktail que dans le bruit blanc. Par contre, on n'observe pas d'effet du type de bruit sur les pics de vitesse de A ni sur ceux du pincement avalé. Les pics de vitesse du mouvement peuvent être reliés à la notion d'effort articulatoire [10].

Table 2 : Évolution des pics de vitesse des mouvements articulatoires dans les deux environnements bruyants.

		Δ dans bb	Δ dans cktl	Δ cktl-Δbb
Vmax de A		+12,7% **	+13,9% ***	+1,2% ns
Vmax de B		+14,9% ***	+29,9% ***	+15,0% ***
Vmax de S		+17,8% ***	+38,5% ***	+20,7% ***
Vmax du Pincement	protrus	-16,5% ***	+3,3% ns	+19,8% **
	avalé	+120,1% **	+127,9% **	+7,8% ns

3.2. Évolution des paramètres acoustiques dans le bruit

Sur les énoncés

La table 3 présente les valeurs moyennes des paramètres acoustiques sur la globalité des énoncés. Les résultats acoustiques obtenus sont cohérents avec les études antérieures sur la parole Lombard [1-3] : l'intensité et la fréquence fondamentale augmentent significativement quel que soit le type de bruit, de même que la durée des énoncés (cf. Table 3). On observe une diminution conséquente de Oq. La différence d'énergie entre le maximum spectral de la zone 80-1500Hz (1ers harmoniques de la voix) et le maximum spectral de la zone 1500-3500Hz augmente significativement dans le bruit. Il en va de même pour la différence d'énergie spectrale entre la zone 80-1500Hz et la zone 3500-5500Hz.

Table 3 : Évolution des paramètres acoustiques dans les deux environnements bruyants.

		Δ dans bb	Δ dans cktl	Δ cktl-Δbb
Intensité		+12,9dB ***	+8,6dB ***	-4,3dB ***
F0		+55,5Hz ***	+65Hz ***	+9,5Hz ***
Énergie spectrale	1500-3500Hz	+23dB ***	+19dB ***	-4dB ***
	3500-5500Hz	+27dB ***	+19dB ***	-8dB ***
Oq		-0,123 ***	-0,118 ***	+0,005 ns
Durée		+11% ***	+8.2% ***	-2,8% *

Le type de bruit a un effet significatif sur l'évolution des paramètres acoustiques, à l'exception de Oq. L'augmentation de l'intensité, de l'énergie spectrale et de la durée des mots est plus importante dans le bruit blanc que dans le bruit cocktail. Par contre, l'élévation de la fréquence fondamentale est plus importante dans le bruit cocktail que dans le bruit blanc.

Sur les voyelles et les consonnes

L'intensité des voyelles augmente davantage dans le bruit que celle des consonnes (cf. Table 4). Par contre, le renforcement de l'énergie spectrale n'est pas significativement différent dans le bruit pour les voyelles et les consonnes sonores du corpus. Les voyelles sont significativement allongées dans le bruit tandis que les consonnes ont tendance à être raccourcies, conformément aux observations de [1-3]. L'augmentation de l'intensité et de l'énergie spectrale dans la zone 3500-5500Hz est plus importante dans le bruit blanc que dans le cocktail, que ce soit sur les voyelles ou les consonnes. Les voyelles sont également plus allongées dans le bruit blanc. Par contre, l'effet du type de bruit sur le renforcement de l'énergie spectrale dans la zone 1500-3500Hz n'est significatif que pour les consonnes, avec un renforcement plus important dans le bruit blanc que dans le bruit cocktail.

Table 4 : Évolution des paramètres acoustiques dans le bruit pour les voyelles et les consonnes.

	Intensité	Énergie spectrale		Durée
		1500-3500Hz	3500-5500Hz	
ΔVoyelles dans bb	+12,9dB ***	+17,7dB ***	+24,1dB ***	25,2% ***
ΔVoyelles dans cktl	+8,4dB ***	+17,4dB ***	+13,3dB ***	16% ***
Différence cktl-bb pour ΔVoyelles	-4,5dB ***	-0,3dB ns	-10,8dB ***	-9,2% ***
ΔConsonnes dans bb	+8,5dB ***	+21,6dB ***	+28,2dB ***	-5,7% ns
ΔConsonnes dans cktl	+3,5dB ***	+19,6dB ***	+11,8dB ***	-3,6% ns
Différence cktl-bb pour ΔConsonnes	-5,0dB ***	-2,0% ***	-16,4dB ***	+2,1% ns
Différence ΔVoyelles - ΔCons	+4,6dB ***	+ 2,9dB ns	+1,3dB ns	+25,2% ***

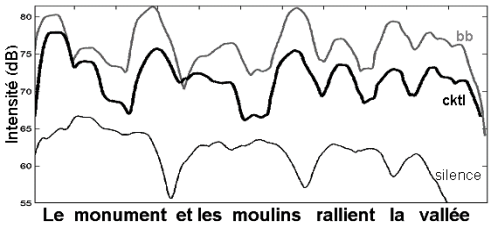


Figure 3 : Intensité d'un énoncé dans le silence, le bruit blanc (bb) et le bruit cocktail (cktl).

Comme attendu, les voyelles sont plus fortes que les consonnes dans le silence (effet de 4,5dB, p<0,001), mais l'intensité suit globalement les constituants SVO, avec une chute marquée au niveau de leur frontière. Dans le bruit, on constate un renforcement de la dynamique des "lobes" d'intensité par syllabe (+4dB, p<0,001), comme illustré Figure 3, pouvant contribuer à une perception plus hâchée de la parole.

Sur les trois voyelles /u/, /i/, et /a/

Nous avons ensuite examiné l'évolution dans le bruit des trois voyelles /u/, /i/, et /a/, formant les extrema du triangle vocalique. Seules l'intensité et l'énergie spectrale évoluent de façon différente dans le bruit pour ces trois voyelles (cf. Table 5). L'évolution de la durée et de Oq ne dépend pas significativement de la catégorie phonétique.

Table 5 : Évolution des paramètres acoustiques dans le bruit pour les 3 voyelles /a/, /i/ et /u/. La dernière colonne du tableau représente la moyenne quadratique des différences entre /u/ et /i/, entre /u/ et /a/ et entre /i/ et /a/.

		Δ / u / dans le bruit	Δ / i / dans le bruit	Δ / a / dans le bruit	Effet du type de voyelle
Intensité		+9,6dB ***	+8,9dB ***	+14,6dB ***	4,5dB ***
Énergie spectrale	1500-3500Hz	+18,4dB ***	+20,5dB ***	+11,4dB ***	7,0dB ***
	3500-5500Hz	+21,5dB ***	+19,2dB ***	+13,8dB ***	5,7dB **
Oq		-0,111 ***	-0,115 ***	-0,092 ***	0,017 ns
Durée		26,7% ***	23,2% ***	14,3% **	9,2% ns

L'observation du triangle vocalique montre une tendance globale d'augmentation du premier formant dans le bruit, en particulier pour la voyelle ouverte /a/ et la voyelle protruse /u/ (cf. Figure4). Ce résultat est en accord avec les observations de Van Summers et al., 1988 [3] et de Rostolland [17]. Le deuxième formant est sensiblement modifié pour le /i/ et le /a/ et augmente considérablement dans le bruit pour le /u/ . L'augmentation de F1 est indépendante de l'augmentation de la F0 (r=0,22 dans le bruit blanc et r=0,014 dans le bruit cocktail).

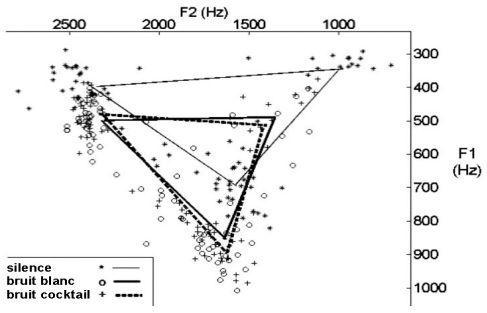


Figure 4 : Évolution du triangle vocalique dans le bruit blanc et le bruit cocktail. (Les sommets des triangles correspondent aux barycentres des /a/, /i/ et /u/ de chaque condition, dans le plan F1/F2).

3.3. Corrélations entre paramètres acoustiques et articulatoires

Sur la globalité des énoncés, on observe une très forte corrélation entre l'amplitude maximale de A et B (r=0,91) ainsi qu'entre les pics de vitesse de A et B (r=0,70). La fréquence fondamentale et l'énergie spectrale sont corrélées avec les maxima de B (r=0,66 et r=0,70) davantage que l'intensité avec les maxima de B (r=0,59). La durée est très peu corrélée avec les paramètres acoustiques et articulatoires (|r|<0,15).

4. DISCUSSION ET CONCLUSION

On observe une hyper-articulation en environnement bruyant, portant davantage sur l'ouverture et l'étirement des lèvres que sur leur pincement. Elle concerne autant l'amplitude des mouvements articulatoires que leurs pics de vitesse, pouvant être reliés à la force articulatoire. On observe un renforcement conjoint des paramètres acoustiques, qui confirment les observations d'études antérieures [1-3].

Le type de bruit a un effet significatif sur les paramètres acoustiques (augmentation plus importante dans le bruit blanc que dans le bruit cocktail, excepté pour la F0). Cette influence du contenu spectral a déjà été évoqué ou mise en évidence dans d'autres études [2,14]. Il a également un effet sur les paramètres articulatoires (augmentation plus importante dans le bruit cocktail que dans le bruit blanc). Par ailleurs, l'énergie spectrale est le paramètre de plus fort taux d'augmentation dans le bruit. L'ouverture des lèvres est davantage corrélée à l'énergie spectrale qu'à l'intensité de la voix. Ces différents points argumentent plutôt en faveur d'une adaptation acoustique et articulatoire dans le bruit cherchant à augmenter l'émergence spectrale de la voix plutôt qu'à augmenter prioritairement le rapport voix sur bruit. On observe également que l'adaptation acoustique favorise les voyelles dans le bruit par rapport aux consonnes. Cela pourrait sembler logique pour des consonnes non sonores, risquant d'être masquées par le bruit environnant. Le fait que l'on observe cette adaptation également pour des consonnes sonores pourrait corroborer l'hypothèse d'une intelligibilité dans le bruit reposant sur des patrons vocaliques, proposée par Dohalska [16]. Il est à noter que des observations inverses ont été faites en parole claire [5]. Enfin, les premières observations articulatoires et formantiques ne semblent pas aller dans le sens d'une spécialisation des voyelles en leur lieu d'articulation, à part pour les voyelles ouvertes. Cela pourrait expliquer l'intelligibilité dégradée de la voix criée observée par Rostolland [17]. Dans le but d'explorer plus rigoureusement l'hyper-articulation des différents types de voyelles, un deuxième corpus contrôlant la coarticulation vient d'être enregistré. Il permettra d'analyser la protrusion en plus des paramètres d'étirement et d'ouverture. En outre, nous avons enregistré plusieurs locuteurs pour vérifier si les résultats de cette présente étude peuvent être généralisés. Il serait également intéressant de prolonger cette exploration à l'analyse des mouvements linguaux pour déterminer s'il existe une compensation au niveau de la langue, notamment pour les voyelles protruses.

5. REMERCIEMENTS

Nous tenons à remercier Jacques Poitevineau, Nathalie Henrich, Christophe Savariaux, Alain Arnal, Aude Noiray, Claire Lalevée et Coriandre Vilain pour leur contribution à cette étude. La participation de Pauline Welby à cette recherche a été soutenue par une bourse internationale Marie Curie du 6th European Community Framework Programme.

6. BIBLIOGRAPHIE

[1] A. Castellanos, J. M. Benedi et F. Casacuberta. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Comm.,* 20 : 23–35, 1996.

[2] J. C. Junqua. The Lombard reflex and it role on human listener and automatic speech recognisers. *JASA,* 93 : 510–524, 1993.

[3] W. Van Summers, D. B. Pisoni, R. H. Bernacki et R. I. Pedlow, M. A. Stokes. Effect of noise on speech production: acoustic and perceptual analyses. *JASA,* 84 : 912–928, 1988.

[4] M. Garnier, N. Henrich, D. Dubois et J. D. Polack. Est-il valide de considérer l'effet Lombard comme un phénomène linéaire en fonction du niveau de bruit ? *Actes du 8ème Congrès Français d'Acoustique, Tours,* 2006.

[5] B. Lindblom. Speech transforms. *Speech Comm*., 11 : 357–368, 1992.

[6] H. Lœvenbruck. Effets articulatoires de l'emphase contrastive sur la Phrase Accentuelle en français. *Actes des XXIIIèmes JEP*, pages 165–168, 2000.

[7] R. Schulman. Articulatory dynamics of loud and normal speech. *JASA,* 85 : 295–312, 1988.

[8] J. C. Junqua. Acoustic and production pilot studies of speech vowels produced in noise. *Actes de ICSLP*, pages 811–814, 1992.

[9] J. Kim, C. Davis, G. Vignali et H. Hill. A visual concomitant of the Lombard reflex. *Actes de Auditory-Visual Speech Processing 2005*, pages 17–21, 2005.

[10] L. Bailly. Étude articulatoire de la parole produite en environnement bruyant. *Mémoire de DEA ATIAM*, *université Paris 6,* 2005.

[11] M. T. Lallouache, Un poste "Visage-Parole" couleur. Acquisition et traitement automatique des contours des lèvres. *Thèse de doctorat, INPG*, 1991.

[12] M. Dohen. Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français. Thèse de doctorat, INPG, 2005.

[13] J. Zeiliger. BD_Bruit, une base de données de parole de locuteurs soumis à du bruit. *Actes des Xèmes JEP,* 287–290, 1994.

[14] S. Ternström, M. Sodersten et M. Bohman, Cancellation of simulated environmental noise as a tool. *J. Voice,* 16 : 195–206, 2002.

[15] W. L. Nelson, Physical principles for economies of skilled movements. *Biol. Cybern.,* 46 : 135–147, 1983.

[16] M. Dohalska et J. Mejvaldova. Rôle de la prosodie dans la communication en milieu bruité. *Actes des Xèmes JEP,* pages 265–268, 2000.

[17] D. Rostolland, Phonetic structure of shouted voice. *Acustica,* 51 : 80–89, 1982.

« Locus equation » pour les consonnes /b/, /d/ et /ɣ/ du vietnamien

Eric Castelli[1] *- Anne Hierholtz*[1,2]

[1]Centre de recherche international MICA
IP Hanoi - CNRS/UMI-2954 - INP Grenoble
1 Dai Cô Viêt, Hanoi, Vietnam
[2]Ecole Polytechnique Fédérale de Lausanne
CH-1015, Lausanne, Suisse
Eric.Castelli@mica.edu.vn, Anne.Hierholtz@mica.edu.vn
http://www.mica.edu.vn

ABSTRACT

Locus equation measurements are one approach used to characterise vocal tract resonances during stop consonant production, the place of articulation of the consonant and the nature of the vowel-consonant transition. Taking up again previous literature studies, the aim of the present work is to applying locus equation measurements to Vietnamese language, for two stop voiced consonants /b/ and /d/, and for the specific velar voiced fricative consonant /ɣ/. Because Vietnamese is a tonal language, a specific corpus was built, in order to avoid as much as possible tonal co-articulation effects. We also take into account the two specific vowels that present dynamic characteristics. Comparisons with other languages are given at the end of this study.

1. INTRODUCTION

L'appareil de production du signal de parole peut être considéré en première approximation comme un système simple source + filtre (ou excitateur + résonateur). Lors de la production des voyelles, la principale source d'excitation est constituée des vibrations des cordes vocales et celle-ci est pseudopériodique. Les spectres présentent alors des pics bien visibles, appelés formants, qui peuvent être considérés comme la signature de la voyelle. Ces formants représentent les résonances du conduit vocal et dépendent de la géométrie de ce même conduit. En revanche, pendant la production des consonnes sourdes, la source d'excitation est essentiellement bruitée, avec une source de bruit qui n'est pas forcément située au début du conduit vocal, et les spectres directement calculés sur le signal présentent une répartition assez uniforme des fréquences. Dans le cas des consonnes occlusives, le signal est même partiellement nul juste avant le bruit de plosion puisque le conduit vocal est complètement fermé. Il est alors difficile, voire impossible, de déterminer les résonances du conduit vocal. Cependant, que cela soit pendant la production des voyelles ou pendant celle des consonnes, le conduit vocal est toujours présent, avec une géométrie présentant des résonances.

L'idée du locus consiste donc à mesurer les transitions formantiques pendant la production de séquences CVC et de prolonger artificiellement les positions des résonances dans les consonnes où il est souvent impossible de les déterminer. La position présumée des trois premières résonances (formants F1, F2 et F3) et la prolongation des trois courbes d'évolution au milieu de la consonne s'appellent « le Locus Consonantique » [1, 2]. Chaque consonne de chaque langue peut ainsi être caractérisée. Cependant, le locus ne renseigne pas seulement sur la géométrie du conduit vocal pendant la production de la consonne mais il apporte des informations pertinentes sur la nature de la transition [3].

Le vietnamien est une langue tonale sur laquelle peu d'études ont été menées. Nous pouvons trouver dans la littérature quelques travaux sur la phonétique de la langue mais ses caractéristiques phonologiques restent à étudier car elles permettront de comprendre ses stratégies de production spécifiques. Quelques propriétés spectrales et acoustiques caractérisant les voyelles et les tons sont connues [4, 5] mais aucune étude sur les consonnes n'a encore été menée à notre connaissance. Nous nous proposons d'appliquer l'équation du locus (*locus equation*) à la langue vietnamienne pour caractériser ses consonnes plosives voisées et sa consonne fricative vélaire. Cette caractérisation, outre d'approfondir nos connaissances de la langue, permettra pour une utilisation future de synthétiseurs anthropomorphiques, de mieux piloter les évolutions géométriques du conduit vocal.

2. LANGUE VIETNAMIENNE

Il est généralement admis que le vietnamien présente 9 voyelles acoustiquement différentiables, représentées par les caractères vietnamiens « i/y, ê, e, o, ô, a/ă, ơ/â, ư, u » correspondant phonétiquement respectivement à /i/, /e/, /ɛ/, /ɔ/, /o/, /a/, /ɣ/, /ɯ/ & /u/. Les deux caractères « ă » et « â » peuvent être considérés comme une autre transcription des voyelles /ɣ/ et /a/, cependant, de récents travaux ont montré que les voyelles correspondantes notées /ă/ et /ɣ̆/ présentent des caractéristiques dynamiques [4]. Il y a 21 consonnes en vietnamien standard avec quelques différences dépendantes des dialectes prononcés. Ainsi dans les régions du nord du Vietnam, aucune distinction n'est faite entre les phonèmes /s/ et /ʂ/, /z/ et /z̻/, et /c/ et /t̻/. Notons au passage que le vietnamien présente seulement deux lieux d'articulation pour les consonnes plosives, qui sont labiales pour /p, b/

et alvéolaires pour /t, d/, alors qu'en français, il existe les consonnes vélaires /k, g/. En effet, les consonnes vélaires du vietnamien /ɣ/ et /χ/, respectivement voisée et non voisée, sont considérées non pas comme des plosives, mais comme des fricatives, la consonne /g/ n'existant pas en vietnamien.

3. EXPERIMENTATION

Les consonnes voisées plosives du vietnamien sont /b/ et /d/, transcrites respectivement par les caractères vietnamiens « b » et « đ », suivies par l'une des 11 voyelles vietnamiennes /a/, /ɛ/, /e/, /i/, /ɔ/, /o/, /u/, /ɣ/, /ɯ/, /ă/ et /ɤ̆/. Chaque locuteur vietnamien prononce les syllabes CV (ou CVC) dans la phrase porteuse suivante : « Nói CV/CVC êm ru » qui veut dire « Prononce CV (ou CVC) doucement ». Plusieurs considérations ont guidé le choix de cette phrase porteuse. Premièrement, le vietnamien étant une langue tonale, pour éviter d'éventuels phénomènes d'interférence avec les tons, nous avons souhaité une phrase prononcée en contexte de tons monotones. La condition idéale pour mesurer les formants dans les voyelles, aurait été de faire suivre la voyelle par une consonne occlusive non voisée /t/ ou /p/, ou bien voisée /d/ ou /b/. Cependant, en vietnamien, les syllabes fermées se terminant par ces consonnes ne se prononcent qu'avec des tons à la dynamique complexe. C'est pourquoi, pour les 9 voyelles normales, nous avons gardé une production de syllabes CV ouvertes, en choisissant le mot suivant dans la phrase tel qu'il commence par une voyelle neutre, pour laquelle nous savons que la géométrie du conduit vocal correspond à un tube de section constante, fermé à la glotte et ouvert aux lèvres. Deuxièmement, nous devons tenir compte du comportement dynamique des voyelles /ă/ et /ɤ̆/ qui n'existent qu'en contexte de syllabes fermées et avec la consonne nasale /n/ comme son final dans le cas des tons monotones. La consonne nasale ne nous permet pas de mesurer correctement les formants de la voyelle, car il est alors difficile de définir le point cible de la voyelle. C'est pourquoi, nous choisissons pour les deux voyelles spéciales « dynamiques », une syllabe CVC avec la consone finale /t/ et le ton montant (habituellement noté ton 5) pour que cette configuration augmente le nombre de mots au contenu sémantique significatif en vietnamien (chaque exemple de syllabe CV ou CVC correspond à un mot ayant alors une signification). Chaque locuteur prononce donc les deux consonnes /b/ et /d/ associées avec les 11 voyelles, ce qui correspond à 22 syllabes différentes.

Nous étendons alors notre étude à la consonne vélaire fricative voisée /ɣ/ transcrite avec les caractères vietnamiens « g » ou « gh » (« g » avec les voyelles /a/ et /u/, « gh » avec les voyelles /ɛ/ ou /i/, « h » pouvant alors être considéré comme une représentation de la propriété fricative de la consonne). Pour cette consonne, la principale difficulté réside dans la détection d'un relâchement équivalent au « burst » des consonnes occlusives pour mesurer la valeur du 2[ème] formant au démarrage (onset) de la voyelle. Chaque sujet prononce des syllabes CV (ou CVC), combinaisons de la consonne fricative vélaire et des 11 voyelles vietnamiennes, ce qui rajoute 11 syllabes aux 22 syllabes précédentes

3.1. Mesures

Chaque séquence de signal de parole est enregistrée 5 fois par 8 locuteurs (4 hommes et 4 femmes) originaires du Nord du Vietnam (vietnamien officiel) avec un ordinateur dans un studio calme et sans bruit (mais qui n'est pas une véritable chambre sourde) à la fréquence d'échantillonnage de 11025 Hz. Une pré-emphase permet de rendre plat le spectre des voyelles. Les fréquences des formants sont mesurées avec le logiciel « Praat », en appliquant un fenêtrage gaussien et en calculant les coefficients LPC avec la méthode de Burg (pour plus de détails, se référer à [7]).

3.2. Résultats

Chaque « locus equation » est générée pour une consonne avec toutes les répétions de chaque contexte vocalique (55 séquences = 5 répétitions x 11 voyelles). La pente de régression linéaire, l'ordonnée à l'origine (*Y-intercept*), le coefficient de corrélation R^2 et la déviation standard (*standard error of estimate*) SE sont regroupés pour les consonnes plosives initiales /b/ et /d/, pour chaque locuteur dans le tableau 1. Seule la locutrice F1 présente des valeurs sensiblement différentes de la moyenne du groupe d'étude (pente plus faible et Y-intercept plus grand).

Un exemple de tracé pour un seul locuteur féminin (F3) est donné par la figure 1. Les figures 2a et 2b montrent les équations du locus pour les deux consonnes plosives pour tous les locuteurs.

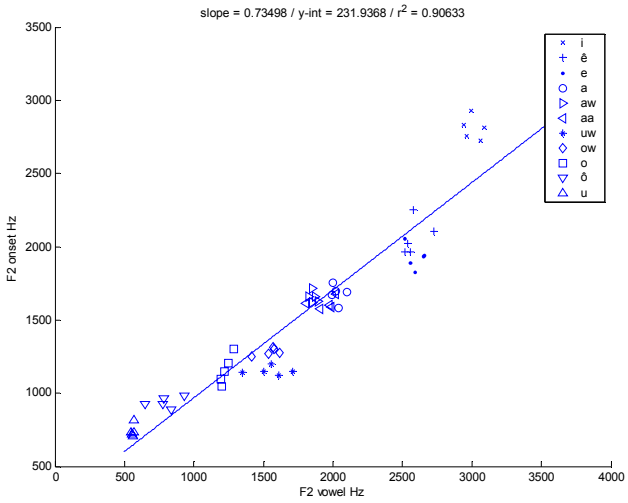


Figure 1 : « Locus equation » de la consonne /b/ (locuteur F3). Les notations « aw, aa, uw et ow » correspondent respectivement aux voyelles ă, â, ơ et ư.

Table 1 : Pente, Y-intercept, R^2 et SE des plosives /b/ et /d/ pour chaque locuteur/.

Locuteur **Fem/Male**	Pente /b/	Pente /d/	Y-intercept Hz /b/	Y-intercept Hz /d/	R^2 /b/	R^2 /d/	SE (Hz) /b/	SE (Hz) /d/
F1 (Hien)	0.36	0.21	680	1603	0.79	0.68	43.66	33.52
F2 (Huong)	0.56	0.29	377	1498	0.86	0.72	52.50	42.42
F3 (Yen)	0.73	0.41	232	1464	0.91	0.84	61.56	48.82
F4 (Hien2)	0.59	0.26	530	1564	0.82	0.66	64.15	44.71
M1 (Hai)	0.65	0.39	244	1103	0.85	0.81	55.47	39.10
M2 (Son)	0.53	0.46	331	979	0.81	0.86	49.28	36.45
M3 (Dat)	0.50	0.31	427	1170	0.84	0.76	43.58	34.32
M4 (Ha)	0.57	0.30	299	1223	0.89	0.67	41.03	40.51
Moyenne	**0.56**	**0.33**	**390**	**1325**	**0.85**	**0.75**	**51**	**40**
Ecart-type	**0.11**	**0.08**	**153**	**235**	**0.04**	**0.08**	**12.01**	**07.77**

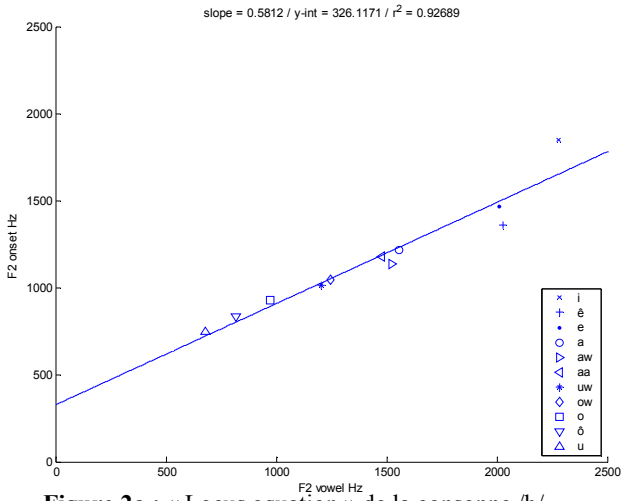


Figure 2a : « Locus equation » de la consonne /b/.

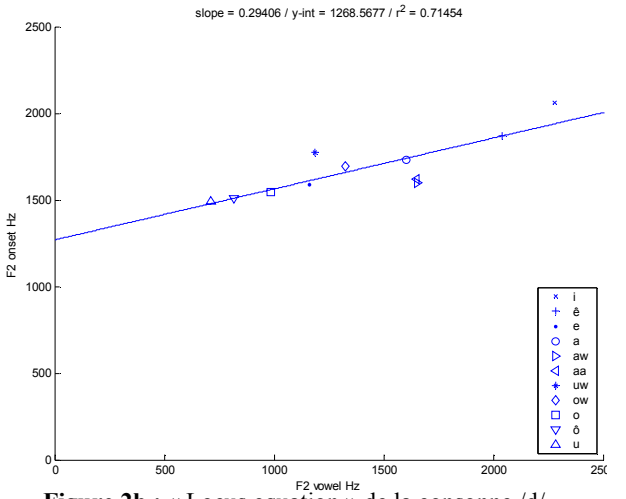


Figure 2b : « Locus equation » de la consonne /d/.

En ce qui concerne la consonne vélaire fricative /ɣ/, la même méthode est appliquée pour caractériser l'équation du locus. Le tableau 2 regroupe les valeurs de pente, « Y-intercept », R^2 et SE. Pour cette consonne aussi, un locuteur (M4) présente des résultats différents de la moyenne des autres locuteurs (pente plus faible et Y-intercept plus grand). Sur la figure 3 est tracé le « locus equation » pour toutes les configurations vocaliques pour tous les sujets.

Table 2 : Pente, Y-intercept, R^2 et SE de la fricative /ɣ/ pour chaque locuteur.

Locuteur **Fem/Male**	Pente /ɣ/	Y-intercept Hz /ɣ/	R^2 /ɣ/	SE (Hz) /ɣ/
F1 (Hien)	0.76	345	0.89	62.25
F2 (Huong)	0.86	331	0.88	73.13
F3 (Yen)	0.85	389	0.95	52.57
F4 (Hien2)	0.66	587	0.76	86.52
M1 (Hai)	0.70	500	0.90	48.13
M2 (Son)	0.75	421	0.91	44.53
M3 (Dat)	0.66	587	0.88	49.84
M4 (Ha)	0.49	689	0.81	48.00
Moyenne	**0.71**	**481**	**0.87**	**58.12**
Ecart-type	**0.12**	**130**	**0.06**	**14.81**

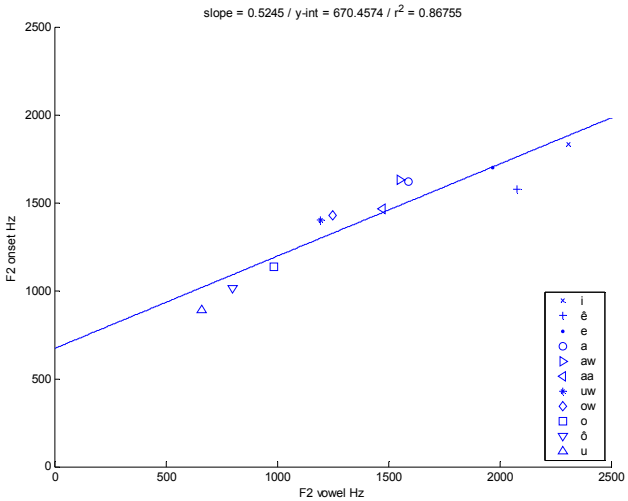


Figure 3 : « Locus equation » de la consonne /ɣ/.

Similairement aux résultats présentés par Sussman et al. [2] pour l'anglais américain pour la consonne plosive /g/, qui présente le même lieu d'articulation que la consonne vietnamienne /ɣ/, deux sous-groupes correspondant aux voyelles antérieures et aux voyelles postérieures (/g/ palatal et /g/ vélaire) peuvent être séparés et caractérisés par deux droites aux pentes différentes.

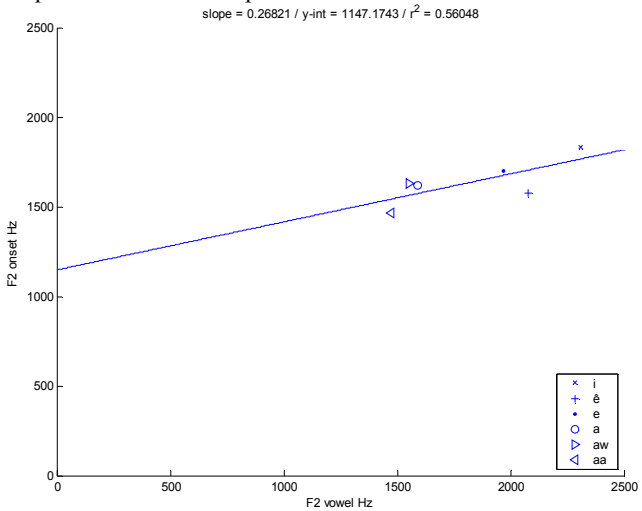


Figure 4a : « Locus equation » de la consonne /ɣ/ pour les voyelles antérieures.

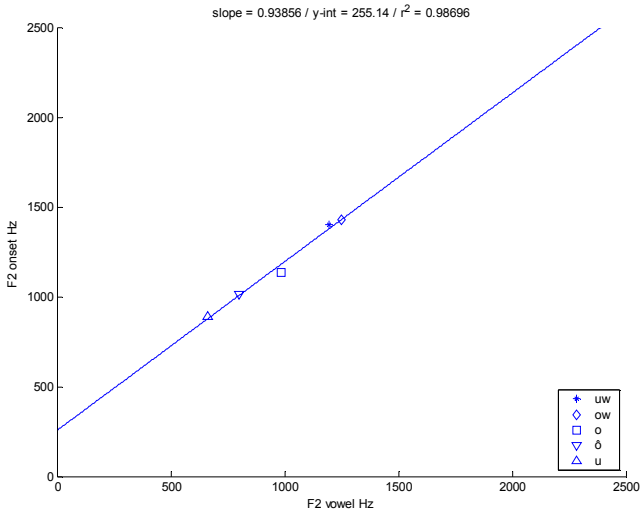


Figure 4b : « Locus equation » de la consonne /ɣ/ pour les voyelles postérieures.

Nous obtenons alors les valeurs pour chaque sous-groupe de R^2 = 0.98, SE = 48Hz pour les voyelles postérieures et de R^2 = 0.56, SE = 180Hz pour les voyelles antérieures. Les figures 4a et 4b tracent les deux locus respectifs.

Comme proposé par Sussman et al. [2], nous utilisons le plan « pente/ordonnée à l'origine » pour représenter les lieux d'articulation des trois consonnes (figure 5). Nous retrouvons bien des zones distinctes qui classifient les deux consonnes /b/ et /d/ et « /ɣ / voyelles postérieures ». Cependant le lieu d'articulation de la consonne /ɣ/ dans le contexte de voyelles antérieures chevauche ceux des consonnes /b/ et /d/.

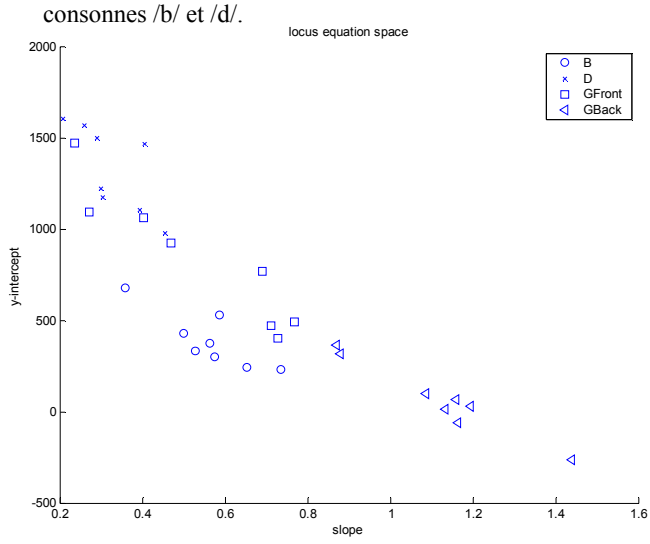


Figure 5 : Lieux d'articulation des consonnes /b/, /d/ et /ɣ/ (dans les deux contextes de voyelles avant et arrières)

4. DISCUSSION

24 équations de locus ont été obtenues, à partir des mesures de 1320 prononciations des consonnes /b/, /d/ et /ɣ/, avec une linéarité excellente et un bon regroupement des points autour de la droite de régression. Sur l'ensemble de nos données nous obtenons un coefficient de corrélation R^2 de 0.82 ce qui est comparable avec les valeurs trouvées dans la littérature (0.86 pour les données de Sussman et al. [2]). Les lieux d'articulation des 2 consonnes plosives et de la consonne vélaire en contexte de voyelles arrières sont bien distincts. Cependant celui de la consonne vélaire palatale semble plus incertain, du fait peut-être de la difficulté de mesurer le formant F2 dans les transitions avec la consonne fricative.

Nous pouvons comparer nos résultats obtenus pour les consonnes vietnamiennes aux données disponibles pour d'autres langues, comme celles de Sussmann et al. [2] sur l'anglais américain (10 locuteurs), de Krull [7] et de Lindblom [1] sur le suédois (respectivement 5 et 1 locuteurs), et de Sussmann et al. [8] sur le thaï, l'arabe et l'urdu (respectivement 6, 3 et 5 locuteurs). Le tableau 3 regroupe les valeurs de pente et d'ordonnée à l'origine et permet de comparer les différentes stratégies d'articulation, qui pour Thaï et VN semblent similaires.

Table 3 : Pente, Y-intercept pour différentes langues

Langues	Pente			Y-intercept Hz		
	/b/	/d/	/g//ɣ/	/b/	/d/	/g//ɣ/
Anglais	0.87	0.43	0.66	106	1073	807
Suédois	0.63	0.32	0.95	487	1096	360
Thaï	0.70	0.30	- - -	228	1425	- - -
Arabe	0.77	0.25	0.92	206	1307	229
Urdu	0.81	0.50	0.97	172	857	212
Vietnamien	0.56	0.33	0.52	390	1325	670

BIBLIOGRAPHIE

[1] Lindblom B. "On vowel reduction" *The Royal Institute of Technology, Speech Transmission Laboratory*, *Stockholm,* Report n°29, 1963

[2] Sussman H., McCaffrey H. A. & Matthews S.A. "An investigation of locus equations as a source of relational invariance for stop place categorization" *J. Acoust. Soc. Am.* 90, 1309-1325, 1991

[3] Chennoukh S., Carré R. & Lindblom B. "Locus equations in the light of articulatory modeling" *J. Acoust. Soc. Am.* 102, 2380-2389, 1997

[4] Castelli E & Carré R. "Production and perception of Vietnamese vowels" *Interspeech-Eurospeech 2005,* Lisbon, Portugal, 2005

[5] Pham Thi N. Y., Castelli E. & Nguyen Q.C. "Gabarits des tons vietnamiens". *JEP 2002* Nancy, pp 23-26, 2002.

[6] Boite R. et al. *"Traitement de la parole"* Presses Polytechniques et Universitaires Romandes, 2000

[7] Krull D. "2nd formant locus patterns & consonant-vowel coarticulation in spontaneous speech" *Phonetic Experimental Research at the Institute of Linguistics,* Univ. of Stockholm, PERILUS VII, 66-70, 1989.

[8] Sussman H., Hoemeke K. & Ahmed F. "A cross-linguistic investigation of locus equations as a relationally invariant descriptor for place of articulation" *J. Acoust. Soc. Am.* 94, 1256-1268, 1993.

Étude de la réduction non linéaire de la dimension du signal de parole

José Anibal Arias, Régine André-Obrecht, Jérôme Farinas et Julien Pinquier

IRIT - Équipe SAMOVA
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, FRANCE
{arias,obrecht,jfarinas,pinquier}@irit.fr
http ://www.irit.fr/recherches/SAMOVA

ABSTRACT

In this article we study some results of the non-linear dimensionality reduction of speech vectors. Spectral clustering, Kernel PCA, Isomap, Laplacian eigenmaps and Locally Linear Embedding are related non-supervised methods that help to discover important caracteristics from data such as high-density regions or low-dimensional surfaces (manifolds). This reduction of dimension is a necessary step when we want to model speech sequences with discriminative/generative functions such as Support Vector Machines or Gaussian Process.

1. INTRODUCTION

Le signal de parole est une source d'information ne pouvant pas être modélisée de façon exhaustive sans prendre en compte sa dimension temporelle. On ne peut se limiter à le considérer comme une variable aléatoire représentée par un vecteur de paramètres extrait en utilisant un fenétrage fixe (centiseconde par exemple). Pour le signal de parole, les modèles de Markov cachés (HMM) ont par exemple permis d'étendre la modélisation statistique par mélange de lois gaussiennes (GMM), en prenant en compte les enchaînements temporels.

Nous souhaitons étudier l'adéquation de méthodes discriminantes telles les machines à vecteur support (SVM) et génératives telles les Processus Gaussiens (GP) avec le signal de parole. Pour pouvoir prendre en compte l'évolution temporelle du signal, il est courant avec les SVM de modéliser des séquences d'observations [12]. La difficulté va alors provenir de la dimension d des vecteurs d'entrée dans le système et de la taille N de l'ensemble des données d'entrainement. En effet, ces deux paramètres crutiaux influent sur la dimension des matrices de traitements internes aux méthodes, il est nécessaire de les contraindre pour pouvoir obtenir des solutions réalisables.

Nous proposons dans cet article d'étudier les variétés ainsi que le regroupement permettent de réduire le nombre des variables pour le signal de parole, dans le but d'appliquer par la suite des modélisations par SVM et GP. Cela devrait introduire des solutions plus robustes (moins sensibles au bruit et aux données aberrantes) et permettre l'analyse visuelle de la structure de l'information.

Dans la section 2 nous présentons synthétiquement plusieurs méthodes que nous avons étudiées, basées sur une décomposition spectrale. Dans la section suivante nous détaillons les expériences menées ainsi que analyse des ces résultats.

2. MÉTHODES SPECTRALES

Les méthodes basées sur une décomposition spectrale estiment de manière non-supervisée les principales fonctions propres d'un opérateur qui dépend d'une densité de données inconnue. On est capable d'utiliser leurs résultats pour généraliser les fonctions propres à des données externes à l'ensemble d'entraînement [3]. L'hypothèse est que, en dépit de la haute complexité du sujet, les sons de la parole sont groupés en variétés non-linéaires de relativement faible dimension liées au processus de production du signal acoustique.

Les algorithmes spectraux d'estimation de variétés s'appuient, pour un ensemble de vecteurs d'entrée $x_{i,\,i=1\ldots N}$, $x_i \in \boldsymbol{R}^d$, sur une matrice de similarité $\overline{K}_{N\times N}$ et conduisent à rechercher ses principaux vecteurs et valeurs propres. La représentation en faible dimension de chaque vecteur x_i en entrée est obtenue en utilisant les j premiers vecteurs propres de $\overline{K}(j \ll d)$. Si l'on veut calculer tenir compte d'un nouveau vecteur x, on utilise la formule de Nyström [5] pour évaluer l'extension des vecteurs propres.

En traitement automatique de la parole, les vecteurs d'entrée sont généralement issus d'une paramétrisation MFCC ou LPC avec d inférieur à la cinquantaine. Les algorithmes spectraux projettent ces données en vecteurs $y_{i,\,i=1\ldots N}$ de dimension très inférieure, idéalement 2 ou 3, pour pouvoir les analyser visuellement.

2.1. Kernel PCA

Schématiquement l'analyse en composantes principales (PCA) est un changement de repère qui vise à privilégier les axes de variance maximale par rapport à un ensemble de données. Les axes où la variance des données est réduite peuvent être éliminés pour atteindre une réduction de la dimensionalité avec une perte minimale d'information. La transformation est, par essence, linéaire (matrice de passage orthogonale). Or, pour les vecteurs de la parole, il est souhaitable de pouvoir atteindre des relations non-linéaires ; la méthode Kernel PCA est une première extension de PCA qui l'envisage.

Kernel PCA réalise une analyse en composantes principales dans l'espace $\boldsymbol{K}$ appelé «espace de caractéristiques» généré à l'aide d'une fonction noyau $k(x_i, x_j)$ tel que $x_i, x_j \in \boldsymbol{R}^d,\ k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ [11]. La transformation non-linéaire $\Phi(x)$ implicite dans la fonction noyau permet au Kernel PCA de trouver un sous-espace qui, plus qu'une réduction de dimensionalité, est le résultat d'un processus d'extraction d'information.

Si X est la matrice de l'ensemble des transformées des données d'apprentissage, dans l'espace des caractéristiques, centrées, la matrice C de covariance de ces transformées vérifie : $NC = X'X$. Si $\overline{K}$ est la matrice « kernel », par définition, $\overline{K} = XX'$.

Il vient que si Xu est vecteur propre de NC associé à la valeur propre λ, X est un vecteur propre de $\overline{K}$ associé à la même valeur propre :

$$\begin{aligned} u &= \lambda^{-1/2} \sum_{i=1}^{N} v_i \Phi(x_i) \\ v &= \frac{Xu}{\sqrt{\lambda}} \end{aligned} \tag{1}$$

En associant de cette manière à tout vecteur propre u_j de NC, le vecteur v_j et la valeur propre λ_j, la projection d'un nouveau vecteur $\Phi(x)$ sur la direction u_j est donnée par :

$$\begin{aligned} P(x)_{u_j} &= u_j' \Phi(x) = \left\langle \sum_{i=1}^{n} \alpha_i^j \Phi(x_i), \Phi(x) \right\rangle \\ &= \sum_{i=1}^{n} \alpha_i^j k(x_i, x) \end{aligned} \tag{2}$$

où $\alpha^j = \lambda_j^{-1/2} v_{j,\, j=1\ldots d'}$.

Une réduction d'information s'obtient en conservant les valeurs et vecteurs propres les plus élevés.

2.2. Isomap

Isomap [13] est une généralisation non-linéaire de l'algorithme d'échelle multidimensionnelle (MDS). MDS permet à partir des distances euclidiennes entre points, de déterminer un système de coordonnées réduit qui préserve les distances. L'idée fondamentale du MDS est la définition d'un produit à partir de la distance entre les vecteurs.

Cette définition nécessite de centrer ces vecteurs [4] : l'expression $x_i \cdot x_j$ dépend non seulement des distances 2-à-2 d_{ij}^2, d_{ki}^2, et d_{kj}^2 mais de toutes les autres distances entre points :

$$x_i \cdot x_j = -\frac{1}{2}(d_{ij}^2 - \frac{1}{n}\sum_{k=1}^{n} d_{kj}^2 - \frac{1}{n}\sum_{l=1}^{n} d_{il}^2 + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} d_{kl}^2) \tag{3}$$

L'étude des vecteurs et valeurs propres de la matrice des produits scalaires permet une réduction de la dimension de l'espace d'observations.

Isomap construit un graphe dont les sommets sont les points et les arêtes les distances entre eux. Un sommet est adjacent à un autre seulement s'ils sont proches. On estime la distance géodésique entre chaque paire de données par la distance la plus courte parcourue sur le graphe (algorithme de Floyd ou Djikstra). On applique MDS à partir de ces distances géodésiques pour obtenir le nouveau système de coordonnées.

2.3. Locally Linear Embedding

L'algorithme LLE [10] modélise une variété comme une union de petits espaces linéaires. Il exploite la géométrie locale des points x_i dans l'espace original pour la reproduire dans un espace de plus faible dimension. Chaque point x_i a un voisinage $N(i)$ et l'idée consiste à exprimer x_i comme une combinaison linéaire de ses voisins $N(i)$ et de construire son image dans le nouvel espace y_i en respectant cette relation.

Les combinaisons linéaires sont obtenues en minimisant l'erreur quadratique globale :

$$\sum_i \| x_i - \sum_{j \in N(i)} W_{ij} x_j \|^2 \tag{4}$$

sous les contraintes $\sum_{j \in N(i)} W_{ij} = 1, \forall i$. Ces contraintes assurent l'invariance par translation des points et de ses voisins. Une procédure de résolution consiste à utiliser le Lagrangien en décomposant le problème en N sous problèmes.

On cherche alors un espace Y de dimension d' ($d' \ll d$) et un ensemble de points $y_{i,\, i=1\ldots N}$, $y_i \in Y$ tels que l'équation suivante soit minimale, W_{ij} étant donné :

$$\sum_i \| y_i - \sum_{j \in N(i)} W_{ij} y_j \|^2 \tag{5}$$

Y doit pouvoir être translaté sans affecter la fonction de coût, donc $\sum_i y_i = 0$. Pour éviter des solutions dégénérées, la covariance de Y est diagonale et $\frac{1}{N}\sum_i y_i y_i' = I$. Cette contrainte lie toutes les variables et le problème d'optimisation ne peut pas être décomposé pour chaque i comme précédemment.

La solution est de la forme :

$$(I - W)'(I - W)Y = \frac{1}{N} Y \Lambda \tag{6}$$

La réduction de la dimension à d' est reliée aux d' plus petites valeurs propres non nulles de $(I - W)'(I - W)$.

2.4. Spectral clustering

Les résultats du regroupement de données basé sur une décomposition spectrale sont proches de ceux subjectivement perçus les humains. Contrairement à l'algorithme des k-means, ils sont capables de trouver des classes de structure non convexe. Ces méthodes utilisent les vecteurs propres d'une matrice dérivée de la distance entre les vecteurs $x_i,\ i = 1 \ldots N$ pour déterminer les groupes.

L'algorithme « spectral clustering » proposé par [8] est une approximation de la solution au problème (NP-complet) de la séparation d'un graphe en k-groupes. À partir d'une matrice d'affinité A non négative et symétrique qui représente les distances entre points, est définie une matrice D diagonale dont la valeur (i, i) est la somme de la ligne i de A.

Les vecteurs propres, associés aux plus grandes valeurs propres du Laplacien $L = D^{-1/2} A D^{-1/2}$, permettent de construire une représentation de faible dimension des points originaux. Une procédure k-means aide à déterminer les classes de données.

2.5. Laplacian Eigenmaps

Cet algorithme [2] est une variante et mélange les méthodes précédentes. Une représentation graphique des données est obtenue en considérant comme nœuds les points x_i et comme poids sur les arètes, W_{ij}, les distances calculés avec un noyau Gaussien ou un noyau de type les k-plus proches voisins. Si D est la matrice diagonale avec éléments $D_{ii} = \sum_j W_{ij}$, la fonction à minimiser est :

$$\sum_{ij} (y_i - y_j)^2 W_{ij} \tag{7}$$

De manière similaire au LLE, la minimisation est forte sous les contraintes d'une projection centrée, de variance unitaire. La solution est trouvée grâce aux plus faibles valeurs propres.

2.6. Comparaison des méthodes

Les algorithmes des méthodes de réduction de dimension se déroulent selon des schémas comparables : des suites d'optimisations et de décompositions spectrales sont effectuées à chaque fois.

MDS approche une matrice de Gram de produits scalaires. Cette matrice possède les mêmes valeurs propres que celle de la matrice de covariance de l'ACP : les sorties des deux procédures sont équivalentes.

Selon [6, 9], Isomap, LLE, Laplacian eigenmaps et spectral clustering peuvent être considérés comme des instances de Kernel PCA où la matrice de Gram a été calculée à partir des graphes pondérés au lieu d'une fonction prédéfinie. Ces noyaux sont dits « dépendants des données ». Les graphes reflètent les relations de voisinage des données d'entrée.

3. EXPÉRIENCES

Nous avons effectué des expériences de réduction non linéaire de la dimensionalité, de regroupement et de déroulement de variétés sur des séquences de parole dans un esprit de « fouille de données ». On a utilisé des séquences de vecteurs MFCC du corpus OGI multilangues [7]. Seul les extraits de parole spontanée ont été traités (sous corpus « story-bt », segments de 45 secondes).

3.1. Visualisation des variétés

La visualisation des variétés associées aux séquences de parole obtenues par Isomap et LLE permettent de distinguer une distribution et un regroupement selon les unités phonétiques.

Sur les figures 1 et 2, chaque y, image d'une donnée x, est étiqueté manuellement de la façon suivante : #=pause, #bn=bruit de fond, c=silence occlusive, 0=occlusives, F=fricatives, N=nasales, Vx=voyelles.

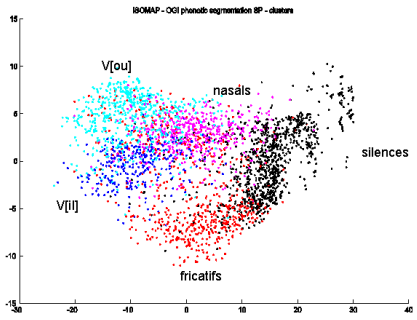


FIG. 1: Représentation d'une séquence de parole obtenue avec l'algorithme Isomap.

Sur la figure 1, on retouve une répartition intéressante des classes. Au milieu, on aperçoit les consonnes avec des regroupements relativement homogènes en silences avant occlusion, nasales, fricatives, occlusives. Et sur la partie gauche de la figure sont regroupées les voyelles.

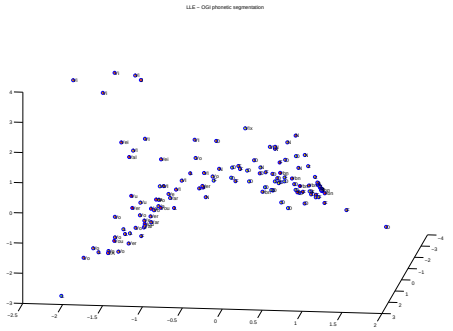


FIG. 2: Représentation d'une séquence de parole obtenue avec la méthode LLE.

Sur la figure 2, les zones de regroupement sont un peu moins homogènes que sur la figure précédente, mais on note tout de même un clivage entre consonnes (à droite) et voyelles (à gauche).

Ces projections permettent d'envisager de réaliser des classifications automatiques. Dans le paragraphe suivant nous allons envisager une telle utilisation.

3.2. Détection automatique de classes

Le deuxième test illustre l'application d'une méthode de classification au sous-espace optenu par un « spectral clustering ». Cette méthode, présentée en [1], permet d'obtenir une classification automatique de régions.

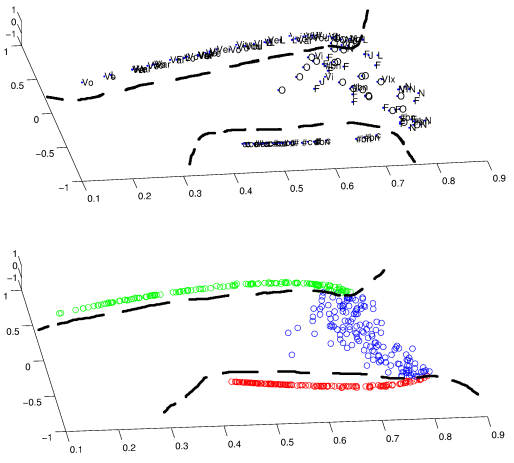


FIG. 3: Détections manuelle et automatique des classes phonétiques par « spectral clustering ». On distingue les trois principales classes phonétiques : silences (en bas), consonnes (au milieu) et voyelles (en haut).

La figure 3, sur la partie haute, est une projection des trois premières dimensions optenu après « spectral clustering », avec affichage des classes avec le même étiquetage que précédemment. La projection en dessous, représente le résultat obtenu après une classification automatique. Les trois figures sont très similaires : la classification automatique nous permet de discriminer les différentes classes.

3.3. Études des variétés

La figure 4 représente les projections sur trois langues de séquences de la parole après réduction non linéaire de la dimension. Seuls les segments consonantiques (détectés automatiquement) sont représentés. Chaque projection laisse apparaitre une forme spécifique à chaque langue. Il est à noter que ces formes apparaissent au bout de quelques secondes : très peu de représentants sont nécessaires pour commencer à les analyser.

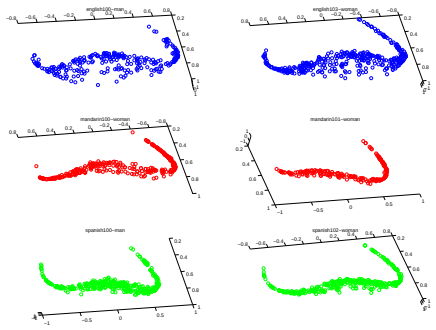


FIG. 4: Répresentation multilangues de séquences de parole par spectral clustering. Les langues étudiées sont l'anglais (en haut), le mandarin (au milieu) et l'espagnol (en bas).

4. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté plusieurs méthodes de réduction non linéaire de séquences de parole, basées sur la décomposition spectrale. Ces méthodes nous ont permis de visualiser en dimension 2 ou 3 l'espace des caractéristiques (section 3.1), en faisant apparaitre des regroupements de grandes classes phonétiques. Dans la section 3.2 nous avons appliqué une classification automatique travaillant dans cet espace projeté, ce qui a permis de les discriminer. La projection de certaines classes phonétiques sur plusieurs langues fait apparaître différentes variétés (formes) associées au processus de production de la parole.

Notre objectif maintenant consiste à découvrir les géométries de ces différentes variétés afin de pouvoir les détecter et les identifier. La modification des contraintes d'optimisation de certaines méthodes, telles LLE ou Isomap, doit nous permettre de reconstituer les formes des distributions des données acoustiques afin de les caractériser et de les comparer.

Nous nous interessons également à la méthode de Nyström [3]. Celle-ci évite de recalculer systématiquement les vecteurs propres des matrices semi-définies positives que l'on a trouvés comme solution avec les méthodes de réduction spectrales non-linéaire en les généralisant aux nouveaux vecteurs d'entrée.

Nous pensons que, appliqués à la parole ces méthodes pourront être utiles pour la caractérisation de certaines modes de production acoustiques, en utilisant des méthodes discriminantes (type SVM) ou génératives (GP) pour des tâches d'indexation sonore, d'identification de langues ou de reconnaissance de locuteur.

RÉFÉRENCES

[1] A. Arias. Unsupervised identification of speech segments using kernel methods for clustering. In *European conference on Speech Communication and Technology*, 2005.

[2] M. Belkin. and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6) :1373–1396, 2003.

[3] Y. Bengio, O. Delalleau, N. Le Roux, and J.F. Paiement. Spectral dimensionality reduction. *Centre interuniversitaire de recherche en analyse des organisations (CIRANO)*, 27, 2004.

[4] I. Borg and P. Groenen. *Modern Multidimensional Scaling : Theory and Applications*. Springer, 1997.

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nyström method. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 26, 2004.

[6] J. Ham, D. Lee, S. Mike, , and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty First International Conference on Machine Learning*, pages 369–376, 2004.

[7] Yeshwant Kumar Muthusamy, Ronald A. Cole, and B. T. Oshika. The ogi multilanguage telephone speech corpus. In *International Conference on Speech and Language Processing*, volume 2, pages 895–898, October 1992.

[8] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

[9] J. Platt. Fast embedding of sparse similarity graphs. *Advances in Neural Information Processing Systems*, 2004.

[10] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(22) :2323–2326, 2000.

[11] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[12] H. Shimodaira, K.I. Noma, M. Nakai, and Sagayama S. Support vector machine with dynamic time-alignment kernel for speech recognition. In *Proc. Interspeech*, 2001.

[13] J.B. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(22) :2319–2322, 2000.

Estimation de la fréquence des formants basée sur une transformée en ondelettes complexes

Laurence Cnockaert[*]*, Jean Schoentgen*[†]*et Francis Grenez*

Université Libre de Bruxelles
Faculté des Sciences Appliquées, Service Ondes et Signaux
Av. F.D. Roosevelt 50, 1050 Bruxelles, Belgique
lcnockae@ulb.ac.be

ABSTRACT

The objective of this paper is to evaluate the performances of a formant estimation method in tracking variations due to the vocal tract movement during the production of sustained vowel. The formant frequency estimation is based on the instantaneous frequency obtained by means of a complex wavelet transform and is synchronised with the glottal cycle. Results for synthetic speech signals show that the precision of the formant frequency estimation is high. However, the estimated results are influenced by variations of the vocal frequency and variations of close formants. The method is illustrated for real speech.

1. INTRODUCTION

L'estimation des caractéristiques du conduit vocal à partir du signal de parole est un domaine de recherche important, notamment à cause de son utilité pour la compréhension et la modélisation du mécanisme de production de la parole. Pour décrire le conduit vocal, on mesure généralement les caractéristiques des formants qui sont les pics observés dans le spectre du signal vocal et qui correspondent aux résonances libres dans le conduit vocal. Le but de cet article est de caractériser les variations involontaires de la position des articulateurs lors de la production de voyelles soutenues, par l'intermédiaire des fréquences des formants. Une technique précise de mesure de la fréquence des formants est nécessaire à cet effet.

Les propriétés du conduit vocal varient dans le temps. D'une part, la forme du conduit vocal varie durant la production de la parole à cause des mouvements des articulateurs. D'autre part, des variations apparaissent au rythme du cycle glottique, à cause de la vibration des cordes vocales [6]. En effet, les cordes vocales oscillent entre une phase fermée et une phase ouverte, ce qui modifie les caractéristiques du système : pendant la phase fermée, le conduit vocal est fermé à la glotte et le signal de parole résulte des résonances libres dans le conduit, tandis que pendant la phase ouverte, le conduit vocal est couplé acoustiquement avec la glotte et la trachée, ce qui modifie les résonances du conduit.

Pour obtenir les meilleures performances dans le suivi des variations temporelles des paramètres des formants, il faut donc que les fenêtres d'analyse aient une longueur effective plus courte que le cycle glottique et soient synchronisées sur celui-ci [7]. Pour étudier les variations du conduit vocal supra-glottique, les caractéristiques pertinentes sont les valeurs des formants pendant la phase fermée de la glotte. Notons que, comme nous nous intéressons aux variations des formants pour des voyelles soutenues, le critère de performance de l'estimation des formants est basé sur la qualité du suivi des mouvements des formants, et non sur la proximité de la fréquence estimée des formants par rapport à la consigne.

Dans cet article, nous étudions une méthode d'estimation non-stationnaire des fréquences des formants, basée sur la fréquence instantanée obtenue au moyen d'une transformée en ondelettes complexes, avec synchronisation des mesures par rapport à la phase fermée du cycle glottique. La performance de la méthode est illustrée sur des signaux de parole synthétiques. L'effet des variations de la fréquence fondamentale et des fréquences des formants sur les estimations des fréquences des formants est étudié. Finalement, quelques résultats sont présentés pour un signal de parole réel.

2. ESTIMATION DES FORMANTS

2.1. Transformée en ondelettes continue

La fréquence instantanée $FI(t)$ d'un signal passe-bande $s(t)$ est généralement définie au moyen de sa transformée de Hilbert $H[s(t)]$[1].

$$\Phi(t) = arg[s(t) + jH[s(t)]] \quad (1)$$

$$FI(t) = \frac{1}{2\pi}\frac{d\Phi(t)}{dt} \quad (2)$$

La transformée en ondelettes continue $CWT(\lambda, t)$ permet également de définir la notion de fréquence instantanée, lorsqu'on utilise une ondelette analytique [4].

La transformée en ondelettes continue d'un signal $x(t)$ est définie comme

$$CWT(\lambda, t) = \int_{-\infty}^{+\infty} x(u)\,\frac{1}{\sqrt{\lambda}}\psi^*\left(\frac{u-t}{\lambda}\right)du, \quad (3)$$

où $\psi(t)$ est l'ondelette-mère, et où $CWT(\lambda, t)$ est le coefficient de la transformée en ondelettes pour un facteur d'échelle λ, à l'instant t.

L'amplitude et la phase des coefficients $CWT(\lambda, t)$ complexes, obtenus à partir d'une ondelette-mère complexe, sont respectivement l'enveloppe et la phase instantanée des composantes spectrales du signal dans la bande de fréquence centrée autour de la fréquence centrale f_c de l'on-

[*]Le premier auteur est boursière du *Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture* (Belgique).

[†]Le deuxième auteur est *Maître de Recherches* du *Fonds National pour la Recherche Scientifique* (Belgique).

delette [5]. La dérivée temporelle de la phase des coefficients $CWT(\lambda, t)$ est donc une estimation de la fréquence instantanée du signal dans cette bande de fréquences. Par conséquent, on peut étudier l'évolution de la fréquence instantanée dans différentes bandes de fréquence du signal au moyen des coefficients de la transformée en ondelettes.

Ici, l'ondelette complexe de Morlet a été utilisée [3] :

$$\psi_{\omega_c}(t) = C\, e^{-i\omega_c t}\left[e^{-\frac{t^2}{2\sigma_t^2}} - \sqrt{2}e^{-\frac{\omega_c^2\sigma_t^2}{4}} e^{-\frac{t^2}{\sigma_t^2}}\right]. \quad (4)$$

L'échelle λ de l'ondelette est déterminée par la fréquence centrale $f_c = \frac{\omega_c}{2\pi}$, qui est la fréquence d'oscillation de l'ondelette. Le produit $\omega_c\sigma_t$ fixe le lien entre la largeur de l'enveloppe gaussienne de l'ondelette et sa fréquence d'oscillation f_c. Pour avoir une famille d'ondelettes, le produit $\omega_c\sigma_t$ doit être constant. Le facteur C normalise l'énergie. La durée effective de l'ondelette peut être définie comme $2\sigma_t$. La forme gaussienne de l'enveloppe de l'ondelette de Morlet minimise le produit des résolutions temporelle et fréquentielle de l'ondelette et permet par conséquent d'optimiser la précision des résultats.

2.2. Application à l'estimation des formants

La transformée en ondelettes continue permet donc de calculer la fréquence instantanée pour différentes bandes de fréquences du signal de parole. Au voisinage des fréquences centrales d'ondelettes dont la cyclicité correspond bien à celle du signal, l'amplitude de la transformée en ondelettes présente un maximum. La fréquence instantanée obtenue à partir de la phase des coefficients de la transformée en ondelettes est alors très proche de la cyclicité du signal et permet d'obtenir la fréquence fondamentale F_0 du signal [2]. De même, pour de plus petites échelles, si la fréquence d'un formant se situe dans la bande passante d'une ondelette, la fréquence instantanée résultante sera très proche de la fréquence du formant. La fréquence du formant obtenue à partir de la fréquence instantanée présente une meilleure résolution fréquentielle que le pas de calcul fréquentiel de la transformée en ondelettes [2]. Elle sera donc utilisée ici.

Pour optimiser les résultats, les fréquences des formants sont préalablement estimées par les racines du polynôme de prédiction LPC du signal. Différentes valeurs de $\omega_c\sigma_t$ sont alors utilisées pour calculer une transformée en ondelette distincte et adaptée autour de l'estimation de chaque formant. Pour le premier formant, on veille à ce que la durée effective des ondelettes soit plus courte qu'un cycle glottique. Pour le deuxième formant, il faut que la résolution fréquentielle soit suffisamment fine pour dissocier le deuxième du premier formant. Pour le troisième formant, on choisit une bande passante de l'ondelette de 800Hz, pour avoir une bonne résolution temporelle tout en étant capable de dissocier le troisième du deuxième formant.

Instants de mesure Obtenir la fréquence instantanée le long des maxima d'amplitude de la transformée en ondelettes ne suffit pas pour obtenir le tracé des fréquences des formants. En effet, la variation au rythme du cycle glottique est encore présente. Il faut donc échantillonner la fréquence instantanée des formants pour en extraire une valeur caractéristique de la phase fermée de chaque cycle glottique. La transformée en ondelettes permet de détecter l'instant de fermeture glottique, qui est caractérisé par un maximum de l'énergie instantanée de la transformée en ondelette. L'énergie instantanée est donc calculée le long du tracé de chaque formant et ses maxima sont détectés. L'instant de mesure est choisi légèrement après le maximum d'énergie (la moitié de la longueur effective de l'ondelette), afin que l'ondelette d'analyse correspondante se situe dans la phase fermée de la glotte.

3. SIMULATIONS SUR DES SIGNAUX SYNTHÉTIQUES

Dans cette section, nous présentons des résultats illustrant le comportement de la méthode d'extraction des formants sur des signaux synthétiques. Le but de ces simulations est de mettre en évidence et de comprendre la précision et les limites de la méthode en décomposant les difficultés.

3.1. Signaux synthétiques

Les signaux synthétiques sont basés sur un modèle source - conduit. Le signal de source est donné par la dérivée temporelle du modèle de débit glottique de Liljencrants et Fant. Le conduit est obtenu par une cascade de filtres IIR du second ordre variables dans le temps, modélisant chacun un formant. Pour modéliser l'interaction source-conduit, la bande passante des formants est modulée de façon synchronisée avec la source. Deux valeurs différentes de bandes passantes caractérisent donc la phase ouverte et la phase fermée de la glotte.

3.2. Résultats

Pour étudier l'influence des paramètres du signal synthétique sur les fréquences de formants mesurées, les cas présentés sont les suivants :
– F_0 fixe, fréquences des formants fixes,
– F_0 fixe, fréquences des formants variables,
– F_0 variable, fréquences des formants fixes.
Les performances sont évaluées sur base de la qualité du suivi des mouvements des formants, et non de la proximité entre la fréquence estimée des formants et la consigne.

Fréquence fondamentale fixe, fréquences des formants fixes Les figures 1 à 3 illustrent l'estimation des fréquences des formants pour un signal synthétique dont la fréquence fondamentale F_0 est de $120Hz$, et les fréquences des formants de $700Hz, 1200Hz$ et $2500Hz$. Les bandes passantes de tous les formants sont de $100Hz$ et de $150Hz$, pour les phases fermées et ouvertes de la glotte.

La figure 1 montre le signal synthétique, ainsi que l'amplitude de sa transformée en ondelettes, pour $\omega_c\sigma_t = 10$. Les étoiles blanches marquent les fréquences estimées des formants. Les pics aux fréquences des formants apparaissent clairement, ainsi que les moment d'excitation où de l'énergie est présente à toutes les fréquences.

La figure 2 montre une coupe de l'amplitude de la transformée en ondelette et de la fréquence instantanée en fonction de la fréquence centrale des ondelettes, pour un instant donné. Les traits pointillés marquent les fréquences de consigne des trois formants. Les lignes verticales montrent les fréquence centrales des ondelettes pour lesquelles il y a un maximum d'amplitude de la transformée en ondelettes. On peut voir les plateaux de la fréquence instantanée aux fréquences des formants, qui permettent d'obtenir la précision fréquentielle des mesures.

Le choix des instants de mesure est illustré à la figure 3. On y voit l'évolution des énergies des trois formants, qui présentent un maximum par cycle glottique, ainsi que l'évolution des formants instantanés. Les losanges marquent les fréquences estimées des formants. L'amplitude de la variation de celles-ci est inférieure à $0.2Hz$. Les variations obtenues pour d'autres valeurs de F_0 et de formants ont le même ordre de grandeur.

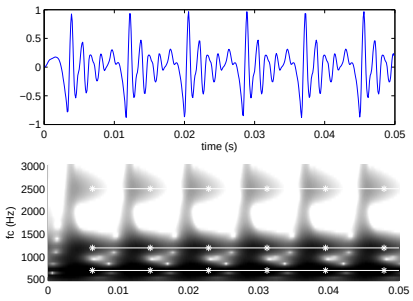


FIG. 1: Signal synthétique et amplitude de sa transformée en ondelettes pour $\omega_c\sigma_t = 10$. Les étoiles blanches marquent les fréquences estimées des formants. Les grandes amplitudes sont représentées en noir, les faibles amplitudes en blanc.

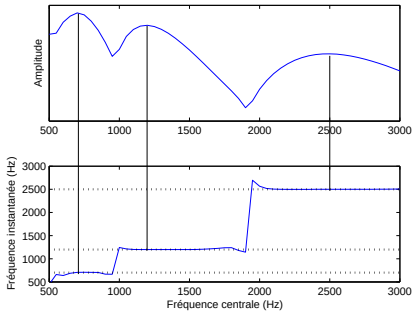


FIG. 2: Coupe de l'amplitude de la transformée en ondelette et de la fréquence instantanée en fonction de la fréquence centrale des ondelettes, pour un instant donné. Les pointillés marquent les fréquences des trois formants. Les lignes verticales montrent les fréquence centrales des ondelettes pour lesquelles il y a un maximum d'amplitude de la transformée en ondelettes.

Fréquence des formants variables Pour tester les performances de la méthode lorsque la fréquence des formants varie, des signaux synthétiques ont été générés avec une fréquence de formant variant linéairement.

Le tableau 1 montre les résultats pour des signaux synthétiques dont la fréquence de F_1 varie, pour deux valeurs de F_0 et deux valeurs de la fréquence de F_2. La fréquence de F_1 varie entre $700Hz$ et $725Hz$, la fréquence de F_2 est de $1100Hz$ ou $1200Hz$, et la fréquence de F_3 est de $2500Hz$. La fréquence fondamentale F_0 est de $100Hz$ ou $125Hz$. Dans le tableau 1, la première partie donne les variations maximales des écarts entre la fréquence mesurée de F_1 et sa consigne. La suite du tableau donne les variations maximales des estimations de F_2 et F_3.

La première partie du tableau 1, montre que, dans tous les

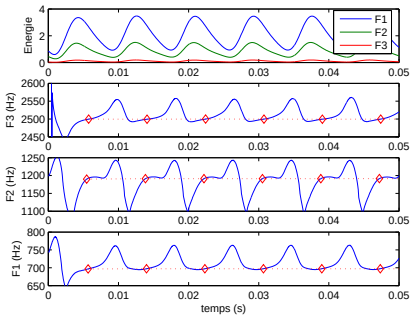


FIG. 3: Evolution de l'énergie des formants et des fréquences instantanées des trois premiers formants. Les losanges marquent les fréquences estimées des formants.

cas, la fréquence estimée de F_1 suit bien la consigne. L'erreur est de l'ordre de $1Hz$, ce qui dépasse légèrement l'erreur obtenue pour un signal à formants fixes. La deuxième partie du tableau 1 montre que la mesure de F_2 est influencée par la variation de F_1, d'autant plus que l'écart fréquentiel entre les deux formants est petit et d'autant plus que F_0 est grande. La mesure de F_3 n'est pas influencé par les variations de F_1.

Les résultats de simulations similaires avec F_2 ou F_3 variable donnent les mêmes conclusions : Premièrement, plus F_0 est élevée, plus l'écart entre le formant variable et la consigne varie. Deuxièmement, plus le formant variable est proche du formant estimé, plus celui-ci varie également.

TAB. 1: Précision de l'estimation des formants pour des signaux synthétiques dont F_1 varie linéairement, pour deux valeurs de F_0 et pour deux valeurs de F_2.

F1 : variation maximale de l'écart entre la mesure et la consigne		
	F0=100Hz	F0=125Hz
F2=1100Hz	0.35Hz	0.56Hz
F2=1200Hz	0.57Hz	1.20Hz
F2 : variation maximale		
	F0=100Hz	F0=125Hz
F2=1100Hz	1.99Hz	8.60Hz
F2=1200Hz	0.27Hz	0.39Hz
F3 : variation maximale		
	F0=100Hz	F0=125Hz
F2=1100Hz	0.004Hz	0.03Hz
F2=1200Hz	0.004Hz	0.03Hz

Fréquence fondamentale variable Pour tester l'effet de la variation de la fréquence fondamentale F_0, des signaux synthétiques ont été générés avec F_0 variant linéairement.

La figure 4 montre l'évolution de F_0 et des fréquences estimées des formants pour un signal synthétique dont F_0 varie linéairement entre $95Hz$ et $105Hz$. Les fréquences de consigne des formants sont de $700Hz$, $1200Hz$ et $2500Hz$. On peut voir que la fréquence des formants n'est pas parfaitement stable et varie en fonction de la proximité de la fréquence du formant avec les harmoniques de F_0. L'effet est plus marqué pour les fréquences de formant plus faibles, mais reste néanmoins inférieur à $2Hz$.

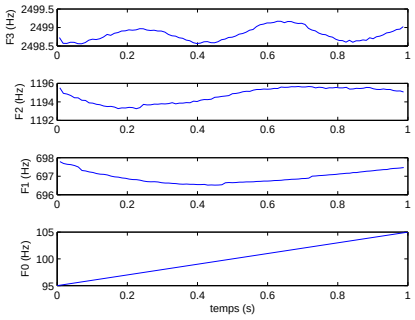


FIG. 4: Evolution de F_0 et des fréquences estimées des trois premiers formants, pour un signal synthétique dont la fréquence fondamentale varie linéairement.

4. APPLICATION À UN SIGNAL RÉEL

La figure 5 montre une voyelle [a] soutenue et l'amplitude de sa transformée en ondelettes avec le paramètre $\omega_c\sigma_t = 10$. Les étoiles correspondent aux fréquences estimées des formants. La figure 6 montre l'énergie instantanée des formants et les fréquences instantanées des trois premiers formants. Les losanges marquent les fréquences estimées des formants. La figure 7 montre la fréquence fondamentale et les fréquences des formants obtenus pour la même voyelle soutenue, pour une durée plus longue. On constate que la méthode permet de détecter et de suivre convenablement les formants.

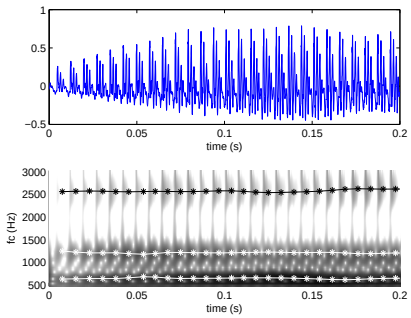


FIG. 5: Signal réel et amplitude de sa transformée en ondelettes pour $\omega_c\sigma_t = 10$. Les étoiles blanches marquent les fréquences estimées des formants. Les grandes amplitudes sont représentées en noir, les faibles en blanc.

5. CONCLUSION

Une méthode d'estimation des fréquences des formants a été proposée. Elle est basée sur la fréquence instantanée obtenue au moyen d'une transformée en ondelettes complexes et est synchronisée par rapport au cycle glottique. Les performances de la méthode d'estimation des formants ont été évaluées pour le suivi des variations dues au mouvement du conduit vocal. Les résultats obtenus sur des signaux synthétiques montrent que la précision de l'estimation de la mesure des formants est très bonne. On constate cependant une influence des variations de la fréquence fondamentale et de la variation des autres formants proches en fréquence. Une validation plus approfondie est en cours sur des signaux synthétiques et réels. L'estimation des variations des formants dues aux mouvements du conduit vocal permettra de comparer les résultats obtenus pour des locuteurs dysphoniques et normophoniques.

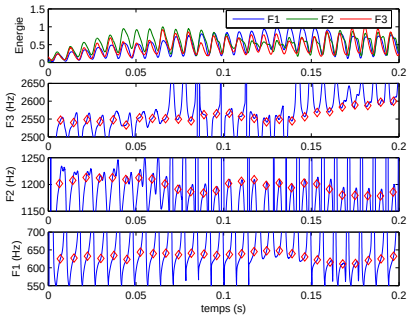


FIG. 6: Evolution de l'énergie des formants et des fréquences instantanées des trois premiers formants. Les losanges marquent les fréquences estimées des formants.

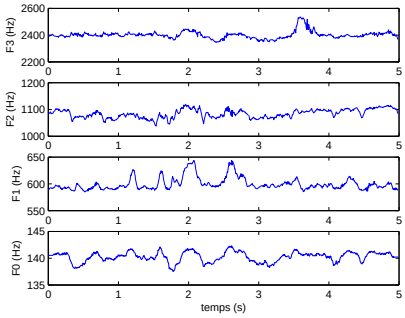


FIG. 7: Evolution de F_0 et des fréquences des trois premiers formants pour un signal réel.

RÉFÉRENCES

[1] B. Boashash. Estimation and interpreting the instantaneous frequency of a signal - part i : Fundamentals. *Proceedings of the IEEE*, 80(4) :520 – 539, 1992.

[2] L. Cnockaert, F. Grenez, and J. Schoentgen. Fundamental frequency estimation and vocal tremor analysis by means of morlet wavelet transforms. *Proc. ICASSP, Philadelphia (USA)*, pages 393–396, 2005.

[3] D.Percival and A.Walden. *Wavelet methods for time series analysis*. Cambridge University Press, 2000.

[4] T. Le-Tien. Some issues of wavelet functions for instantaneous frequency extraction in speech signals. *Proc. IEEE Tencon 1997*, pages 31–34, 1997.

[5] St. Mallat. *A Wavelet Tour of Signal Processing*. San Diego : Academic Press, 2nd edition, 1999.

[6] Pr. Rao and A. D. Barman. Speech formant frequency estimation : evaluating a nonstationary method. *Signal Processing*, 80(8) :1655–1667, 2000.

[7] B. Yegnanarayana and R.N.J. Veldhuis. Extraction of vocal tract system characteristics from speech signals. *IEEE trans. on speech and audio processing*, 6(4) :313–327, july 1998.

Un détecteur d'activité vocale visuel pour résoudre le problème des permutations en séparation de source de parole dans un mélange convolutif

Bertrand Rivet[1,2], *Christine Servière*[2], *Laurent Girin*[1], *Dinh-Tuan Pham*[3], *Christian Jutten*[2]

[1] Institut de la Communication Parlée (ICP)
CNRS UMR 5009, Institut National Polytechnique, Université Stendhal, Grenoble, France
[2] Laboratoire des Images et des Signaux (LIS)
CNRS UMR 5083, Institut National Polytechnique, Université Joseph Fourier, Grenoble, France
[3] Laboratoire de Modélisation et Calcul (LMC)
CNRS UMR 5523, Institut National Polytechnique, Université Joseph Fourier, Grenoble, France

ABSTRACT

Audio-visual speech source separation consists in mixing visual speech processing techniques (e.g. lip parameters tracking) with source separation methods to improve the extraction of a speech signal from a mixture of acoustic signals. In this paper, we present a new method that combines visual information with a separation method based on the sparseness of speech : visual information is used as a voice activity detector which is plugged on an acoustic separation technique. Results show the efficiency of the approach in the difficult case of realistic convolutive mixtures. Moreover, the overall process is quite simpler than previously proposed audiovisual separation schemes.

1. INTRODUCTION

La séparation de source aveugle consiste à retrouver des signaux sources à partir de mélanges de ces signaux, sans connaissances sur la nature du mélange ou sur les sources elles-mêmes. Pour les signaux de parole, la séparation n'est pas complètement aveugle car elle peut s'appuyer sur des propriétés spécifiques de ce signal. Par exemple, leur non-stationarité a été exploitée dans [4, 7]. Cependant, la séparation est encore une tâche difficile, notamment dans le cas où moins de capteurs que de sources sont disponibles, et aussi à cause des indéterminations de permutations et de gains : les signaux de sortie ne peuvent être correctement estimés qu'à un gain près et à une permutation près sur les canaux de sortie [1].

La séparation de source de parole audiovisuelle (SSAV) est un champ de recherche récent intéressant pour résoudre le problème de séparation dans le cas de signaux de parole [2, 8, 10]. Elle consiste à exploiter la bimodalité (audio/visuelle) de la parole pour améliorer les performances des systèmes de séparation acoustiques. En effet, les signaux visuels de la parole, en particulier les mouvements des lèvres du locuteur, fournissent une information complémentaire quand les signaux acoustiques sont dégradés par l'environnement. Sur cette base, Sodoyer *et al.* [8] puis Wang *et al.* [10] ont respectivement proposé d'utiliser un modèle statistique des cohérences entre traits visuels et acoustiques des signaux de parole pour extraire un signal de parole de mélanges de type additif et convolutif. Récemment, Rivet *et al.* [6] ont proposé une approche audiovisuelle similaire pour résoudre à la fois le problème du gain et des permutations dans le cas d'un mélange convolutif.

Dans cette nouvelle étude, on propose une approche différente pour résoudre le problème des permutations, potentiellement plus simple et plus efficace. L'information visuelle de la parole est utilisée comme un détecteur d'activité vocale (DAV) : son rôle est d'attester de la présence ou de l'absence du locuteur correspondant (celui qui est filmé) dans le mélange. Une telle information permet l'extraction du signal émis par ce locuteur.

Ce papier est organisé de la façon suivante. La Section 2 présente les bases du DAV visuel. La Section 3 rappelle les principes de la séparation de source pour des mélanges convolutifs et explique comment le DAV visuel peut être utilisé pour résoudre le problème des permutations pour le locuteur considéré. La Section 4 présente des résultats d'expérimentations.

2. DÉTECTEUR D'ACTIVITÉ VOCALE VISUEL

L'idée centrale du détecteur d'activité vocale visuel (DAV-V) est qu'en général, durant la production de parole, les lèvres bougent, alors qu'elles ne bougent pas (ou beaucoup moins) durant les silences. Nous utilisons le paramètre vidéo suivant :

$$v(m) = \left|\frac{\partial A(m)}{\partial m}\right| + \left|\frac{\partial B(m)}{\partial m}\right| \quad (1)$$

où $A(m)$ et $B(m)$ sont les largeur et hauteur internes du contour labial. Ces paramètres sont extraits automatiquement toutes les 20ms (soit la longueur d'une *trame*) de façon synchrone à l'audio (échantillonné à 16kHz) en utilisant le système d'extraction développé à l'ICP [3]. La classification silence/parole est basée sur un seuillage. Cependant, le seuillage direct de $v(m)$ ne s'avère pas très performant : par exemple, les lèvres peuvent être immobiles pendant plusieurs trames, alors que le locuteur est en train de parler. C'est pourquoi, $v(m)$ est d'abord lissé par intégration temporelle sur T trames consécutives : $V(m) = \sum_{l=0}^{T-1} a_l\, v(m-l)$ où les a_l sont les coefficients d'un filtre passe-bas IIR du premier ordre. La trame m est alors classifiée comme silence si $V(m)$ est inférieure à un seuil δ et elle est classifiée comme parole sinon. Comme expliqué à la Section 3, l'objectif du DAV-V est en fait de détecter les trames de signal où le locuteur filmé *ne produit pas de son*. Pour diminuer le taux de fausses alarmes (décision silence pendant l'activité de parole), seules les séquences d'au moins L trames de silence sont finalement considérées comme silence. Au final, le DAV-V proposé est robuste à n'importe quel bruit environnant et peut être exploité même dans un environnement sonore hautement non-stationnaire, quels que soient le nombre et la nature des sources concurrentes. On peut trouver plus de détails dans [9].

3. SÉPARATION DE SOURCE AVEC DAV-V

Dans cette section, on présente d'abord brièvement le cadre général de la séparation de sources pour des mélanges convolutifs stationaires, puis nous expliquons comment le DAV-V peut être utilisé pour résoudre le problème des permutations.

3.1. Séparation de source de mélange convolutif

Considérons le cas général de N sources $\mathbf{s}(m) = [s_1(m), \cdots, s_N(m)]^T$ à extraire à partir de P observations $\mathbf{x}(m) = [x_1(m), \cdots, x_P(m)]^T$ (^{T} dénote la transposition) : $x_p(m) = \sum_{n=1}^{N} h_{p,n}(m) * s_n(m)$. Les filtres $h_{p,n}(m)$ modélisant la réponse impulsionnelle entre chaque source $s_n(m)$ et le $p^{ème}$ capteur, sont les éléments de la matrice de mélange $H(m)$. Le but de la séparation est de récupérer les sources par un filtrage dual : $\hat{s}_n(m) = \sum_{p=1}^{P} g_{n,p}(m) * x_p(m)$ où les $g_{n,p}(m)$ sont les éléments de la matrice de séparation $G(m)$ et sont estimés de façon à ce que les sources estimées en sortie $\hat{\mathbf{s}}(m) = [\hat{s}_1(m), \cdots, \hat{s}_N(m)]^T$ soient les plus indépendantes possibles (ou au moins décorrélées) deux à deux. Ce problème est généralement traité dans le domaine fréquentiel, par exemple [4, 7], on a alors :

$$X_p(m,f) = \sum_{n=1}^{N} H_{p,n}(f) S_n(m,f) \quad (2)$$

$$\hat{S}_n(m,f) = \sum_{p=1}^{P} G_{n,p}(f) X_p(m,f) \quad (3)$$

où $S_n(m,f)$, $X_p(m,f)$ et $\hat{S}_n(m,f)$ sont respectivement les transformées de Fourier à court terme (TFCT) de $s_n(m)$, $x_p(m)$ et $\hat{s}_n(m)$. $H_{p,n}(f)$ et $G_{n,p}(f)$ sont respectivement les réponses en fréquence des filtres de mélange et de séparation. Des manipulations algébriques simples sur (2) et (3) conduisent à :

$$\Gamma_x(m,f) = H(f)\Gamma_s(m,f)H^H(f) \quad (4)$$

$$\Gamma_{\hat{s}}(m,f) = G(f)\Gamma_x(m,f)G^H(f) \quad (5)$$

où $\Gamma_y(m,f)$ dénote la matrice de densité spectrale de puissance (DSP) à court terme d'un signal multidimensionnel $\mathbf{y}(m)$. $H(f)$ et $G(f)$ sont respectivement les matrices de réponse en fréquence associées aux matrices de mélange et de séparation (^{H} dénote le transposé conjugué). Si on suppose que les sources sont mutuellement indépendantes (ou au moins décorrélées), $\Gamma_s(m,f)$ est diagonale et une séparation efficace doit conduire à une matrice diagonale $\Gamma_{\hat{s}}(m,f)$. Par conséquent, un critère basique pour la séparation est de calculer $\Gamma_x(m,f)$ à partir des observations et d'ajuster la matrice $G(f)$ de telle façon que $\Gamma_{\hat{s}}(m,f)$ soit aussi diagonale que possible. Comme cette condition doit être vérifiée pour n'importe quel indice temporel m, ceci peut être fait par un algorithme de diagonalisation conjointe (*i.e.* la meilleure diagonalisation simultanée de plusieurs matrices) [5], et par la suite nous utilisons l'algorithme de séparation par diagonalisation conjointe des matrices de DSP de Servière et Pham [7].

3.2. Résolution du problème des permutations

La limitation classique des algorithmes de séparation est que pour chaque canal fréquentiel, $G(f)$ ne peut être estimée qu'à un gain près et à une permutation près entre les sources : $G(f) = P(f)\,D(f)\,\hat{H}^{-1}(f)$ où $P(f)$ et $D(f)$ sont une matrice de permutation et une matrice diagonale arbitraires. Plusieurs approches purement audio au problème des permutations ont été proposées (par exemple [4, 7]). Dans [6], nous avons proposé d'utiliser un modèle statistique des cohérences audiovisuelle des signaux de parole pour lever les indéterminations de permutation et de gain . Bien qu'efficace, la méthode à les désavantages de nécessiter un apprentissage hors-ligne et d'être coûteuse en calcul.

Dans cette nouvelle étude, nous simplifions cette approche en exploitant l'information plus simple délivrée par le DAV-V focalisant sur les lèvres du locuteur dont on veut extraire le signal de parole. Le modèle audiovisuel de [6] est remplacée par le DAV-V de la Section 2 et la détection de l'absence de la source d'intérêt permet de régulariser le problème de permutation pour cette source. En effet, pour chaque fréquence f, l'algorithme de séparation fournit une matrice de séparation $G(f)$ qui conduit à une matrice de DSP des sources estimées $\Gamma_{\hat{s}}(m,f)$ diagonale. Le $k^{ème}$ élément de la diagonale de $\Gamma_{\hat{s}}(m,f)$ représente la variation de l'énergie spectrale de la $k^{ème}$ source estimée à la fréquence f au cours du temps m. Appelons le logarithme de cette valeur un *profil* et notons-le $E(f,m;k)$. Notons $\mathcal{T}$ l'ensemble de tous les indices temporels. Supposons maintenant qu'un DAV-V, associé à une source particulière, disons $s_1(m)$, nous fournit l'ensemble des indices temporels $\mathcal{T}_1$ où cette source disparaît du mélange ($\mathcal{T}_1 \subset \mathcal{T}$). Alors le profil $E(f,m;\cdot)$, avec $m \in \mathcal{T}_1$, correspondant à l'estimation de $s_1(m)$ doit être proche de -∞. Par conséquent, nous proposons la technique de régularisation de permutation suivante pour extraire la source particulière $s_1(m)$ du mélange $\mathbf{x}(m)$. A la sortie de l'algorithme de diagonalisation conjointe, on calcule les profils centrés $E_{\mathcal{T}_1}(f;k)$ pendant que $s_1(m)$ est détectée absente, soit pendant $m \in \mathcal{T}_1$:

$$E_{\mathcal{T}_1}(f;k) = \frac{1}{|\mathcal{T}_1|}\sum_{m\in\mathcal{T}_1} E(f,m;k) - \frac{1}{|\mathcal{T}|}\sum_{m\in\mathcal{T}} E(f,m;k) \quad (6)$$

où $|\mathcal{T}_1|$ est le cardinal de l'ensemble $\mathcal{T}_1$. Le centrage permet d'éliminer l'influence du gain non contrôlé, puisque celui-ci devient une constante additive en échelle log. Puis, à partir du fait que le profil centré $E_{\mathcal{T}_1}(f;\cdot)$ correspondant à l'estimation de $s_1(m)$ doit être proche de -∞, pour toutes les fréquences f, on recherche le profil centré de plus faible valeur. On règle alors $P(f)$ de façon à ce que cette valeur minimale corresponde à $E_{\mathcal{T}_1}(f;1)$. L'application de cet ensemble de matrices de permutation $P(f)$ aux matrices de séparation $G(f)$ permet de reconstruire la source $s_1(m)$ sans permutations en fréquence.

Notons que la méthode proposée permet de résoudre les permutations de fréquence pour la source à laquelle le DAV-V est associé, mais il peut rester des permutations entre les autres sources sans conséquence pour l'extraction de $s_1(m)$. Pour extraire plus d'une source, il est nécessaire d'avoir des DAV-V supplémentaires correspondant.

4. EXPÉRIMENTATIONS

Dans cette section, on considère le cas de deux sources mélangées par des matrices de filtres 2×2. Ces filtres sont des filtres RIF de 512 coefficients avec trois échos principaux. Ils sont extraits d'une librairie de réponses impulsionnelles mesurées dans une grande pièce de 3.5m×7m×3m (on peut les trouver

à `http ://sound.medi.mit.edu/ica-bench`). Le corpus utilisé pour la source $s_1(t)$ à extraire est de la parole continue produite par un locuteur masculin enregistré en condition de dialogue spontané. La seconde source est de la parole continue produite par un autre locuteur.

Pour caractériser les performances de l'extraction de $s_1(t)$, on définit un indice de performance :

$$r_1(f) = |GH_{12}(f)/GH_{11}(f)|\,, \qquad (7)$$

où $GH_{i,j}(f)$ est le $(i,j)^{ème}$ élément du système global $GH(f) = G(f)\,H(f)$. Pour une bonne séparation, cet indice doit être proche de 0 (∞ si une permutation a eu lieu).

La Figure 1 présente un exemple de séparation. Les Figures 1(a) et 1(b) montrent les deux sources et les mélanges. Dans ces expériences, dix secondes de signal sont utilisées pour estimer des filtres de séparation de 4096 coefficients (ce qui est donc la taille de toutes les TFCT). Les Figures 1(c), 1(e) et 1(g) montrent les sources estimées dans différentes conditions (voir ci-dessous). Les indices $r_1(f)$ correspondants, tronqués à 1, sont représentés sur les Figures 1(d), 1(f) et 1(h). Sur les Figures 1(a), 1(c), 1(e) et 1(g) les pointillés représentent une indexation manuelle des silences et les traits discontinus représentent la détection automatique obtenue avec le DAV-V de la Section 2. On peut voir que cette détection est performante. De résultats plus détaillés sont donnés dans [9].

Dans la première expérience (Figures 1(c) et 1(d)), les sources sont estimées par l'algorithme de diagonalisation conjointe sans régularisation des permutations. On peut voir que plusieurs blocs de fréquences consécutives sont permutés, ainsi que plusieurs fréquences isolées (Figure 1(d)). Par conséquent, les signaux séparés contiennent des composantes basses/hautes fréquences permutées entre les deux sources (Figure 1(c)). Dans la deuxième expérience (Figures 1(e) et 1(f)), les sources sont estimées par l'algorithme de diagonalisation conjointe utilisant en plus la variation relative des profils [7] : les permutations sont détectées en se basant sur le fait que les profils $E(f,m;k)$ d'une source donnée varient de façon lisse avec la fréquence. On peut voir que les permutations principales sont corrigées, ce qui permet une bonne estimation des sources. Cependant, plusieurs permutations isolées restent présentes bien qu'elles aient une influence limitée sur la qualité de la séparation : sur la Figure 1(f), $(G * H)_{1,1}(n)$ est largement supérieur à $(G * H)_{1,2}(n)$. Dans la dernière expérience (Figure 1(g) et 1(h)), les sources sont estimées en utilisant le DAV-V du locuteur 1 pour détecter les silences de $s_1(m)$ et ainsi régulariser les permutations en utilisant la technique de la Section 3.2. A partir des résultats de [9], on a choisi $\alpha_l = 0.82^l$ et le nombre minimal de trames de silence consécutives L est de 20 (*i.e.* la longueur minimum d'un silence est de 400ms). On peut voir que la méthode proposée permet une très bonne estimation des sources. Il reste quelques permutations isolées mais une investigation plus profonde révèle qu'elles correspondent à des régions des spectres avec une très faible énergie pour les deux sources : elles ont donc une influence très faible sur la qualité de la séparation, comme on peut le voir à la Figure 1(h). Les Figures 1(i) et 1(j) montrent les profils centrés des deux sources estimées avant et après la correction de permutations par le DAV-V. On voit que les blocs permutés sont bien détectés par les profils calculés à partir de la détection de silence : après la régularisation de permutations, on a $E_{\mathcal{T}_1}(f;1) \leq E_{\mathcal{T}_1}(f;2)$ ce qui conduit à une bonne estimation des sources. Ces observations sont confirmées par l'écoute des signaux.

5. CONCLUSION

La détection visuelle, robuste à tout type d'environnement sonore, s'est révélée efficace pour régulariser le problème des permutations selon un principe très simple. Notons aussi que cette méthode visuelle a un avantage important par rapport aux méthodes de séparation purement audio [7] qui fournissent les sources dans un ordre arbitraire (*i.e.* à une permutation globale près sur les sortie même si les permutations sur les différentes fréquences sont corrigées) : l'information visuelle permet d'associer un canal de sortie au locuteur filmé. Dans ce travail, toutes les détections sont réalisées hors-ligne, c'est-à-dire sur des sections de signal relativement larges (de l'ordre de 10s). Nos travaux futurs concerneront le développement d'une version pseudo-temps-réel où les traitements sont effectués en ligne, pour se rapprocher des conditions d'utilisation réelles.

RÉFÉRENCES

[1] Jean-François Cardoso. Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10) :2009–2025, October 1998.

[2] R.M. Dansereau. Co-channel audiovisual speech separation using spectral matching constraints. In *Proc. ICASSP*, Montréal, Canada, 2004.

[3] T. Lallouache. Un poste visage-parole. Acquisition et traitement des contours labiaux. In *Proc. Journées d'Etude sur la Parole (JEP) (French)*, Montréal, 1990.

[4] Lucas Para and Clay Spence. Convolutive blind separation of non stationary sources. *IEEE Trans. Speech Audio Processing*, 8(3) :320–327, May 2000.

[5] Dinh-Tuan Pham. Joint approximate diagonalization of positive definite matrices. *SIAM J. Matrix Anal. And Appl.*, 22(4) :1136–1152, 2001.

[6] Bertrand Rivet, Laurent Girin, and Christian Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Trans. Speech Audio Processing*, (Accepted for publication).

[7] Christine Servière and Dinh-Tuan Pham. A novel method for permutation correction in frequency-domain in blind separation of speech mixtures. In *Proc. ICA*, pages 807–815, Granada, Spain, 2004.

[8] David Sodoyer, Laurent Girin, Christian Jutten, and Jean-Luc Schwartz. Developing an audio-visual speech source separation algorithm. *Speech Comm.*, 44(1–4) :113–125, October 2004.

[9] David Sodoyer, Bertrand Rivet, Laurent Girin, Jean-Luc Schwartz, and Christian Jutten. An analysis of visual speech information applied to voice activity detection. In *Proc. ICASSP*, Toulouse, France, 2006 (accepted).

[10] Wenwu Wang, Darren Cosker, Yulia Hicks, Saied Sanei, and Jonathon A. Chambers. Video assisted speech source separation. In *Proc. ICASSP*, Philadelphia, USA, March 2005.

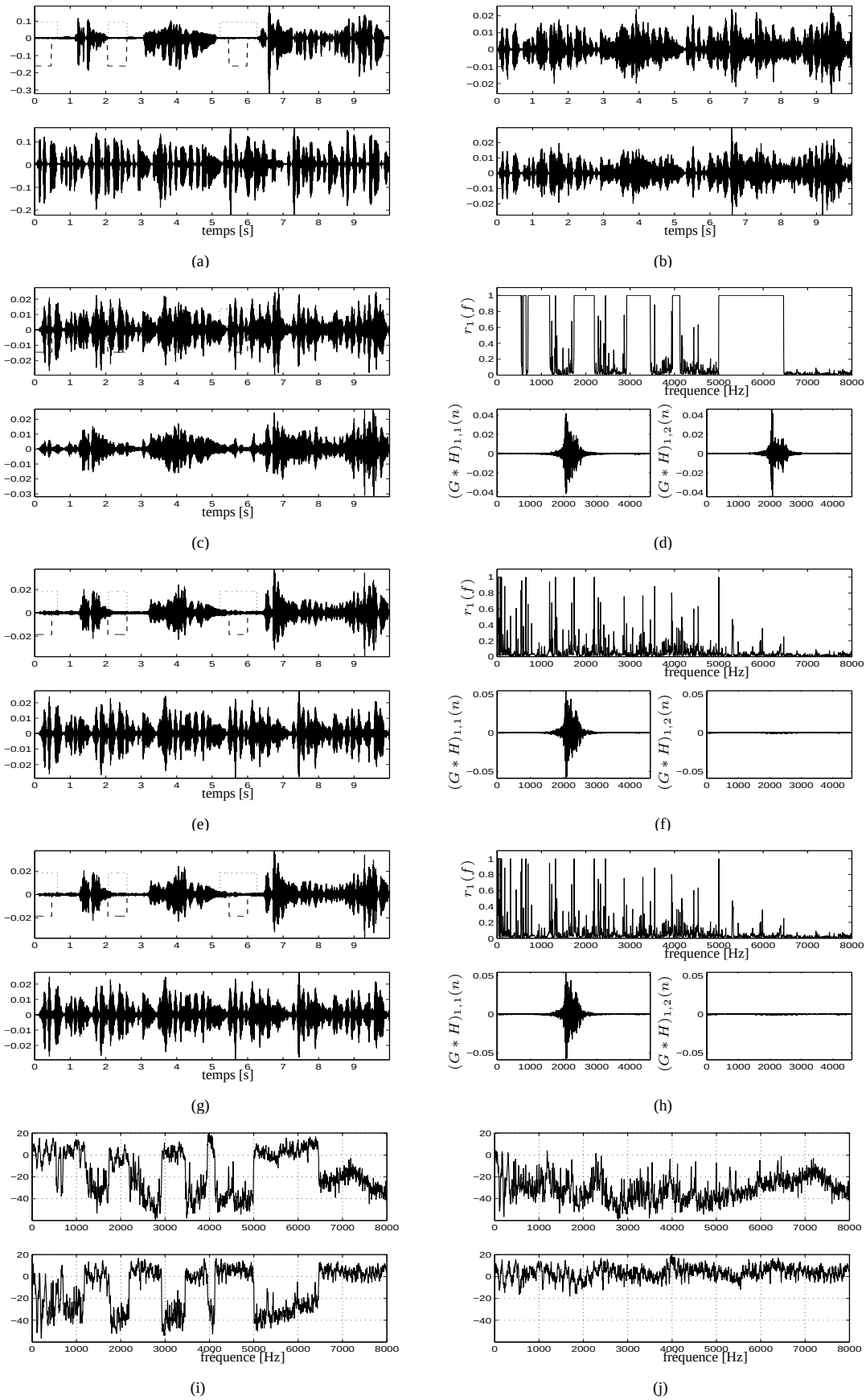


FIG. 1: Sources (a), mélanges (b), sources estimées (c), (e), (g), indice de performance (tronqués à 1) et réponses impulsionnelles du système global $(G*H)(n)$ (d), (f), (h) et profils centrés avant (i) et après (j) corrections des permutations.

Étude de la structure formantique des voyelles produites par des locuteurs bègues en vitesses d'élocution normale et rapide

Fabrice Hirsch, Véronique Ferbach-Hecker, Florence Fauvet & Béatrice Vaxelaire

Institut de Phonétique de Strasbourg – E.A. 1339 – LiLPa -Composante Parole et Cognition - Université Marc Bloch
22, rue Descartes - 67084 Strasbourg
Tél : ++33(0) 88.41.73.64 - Mél : fabrice_hirsch@yahoo.fr

Abstract

The aim of this study is to analyse the steady—state portion of the first two formants (F1) and (F2) in the production of [pVp] sequences, containing vowels [i, a, u] pronounced in two speech rates (normal and fast) by groups of untreated and treated stutterers, and nonstutterer controls. Comparing data between treated or untreated stutterers and controls, a reduction of vowel space is observed for stutterers in a normal speaking rate. When speech rate increases, no reduction of vowel space is noticeable for untreated stutterers, contrary to treated stutterers and controls.

1. Introduction

L'objectif de cette investigation est d'étudier la structure formantique des voyelles produites par des bègues et d'anciens bègues, en comparant leurs données à celles de sujets de contrôle. En outre, nous étudierons également les effets de l'accélération de la vitesse d'élocution sur la structure formantique des voyelles, et cela pour les trois groupes de locuteurs.

Nous avons choisi d'analyser les voyelles orales [i, a, u] du français, à partir de données spectrographiques, dans deux conditions de vitesse d'élocution : normale et rapide. Le choix de ces trois voyelles est motivé par le fait qu'elles représentent les extrêmes du triangle vocalique du français. En les étudiant, il devient ainsi possible d'explorer les limites de l'espace vocalique maximal, puisque leurs productions reflèteraient les capacités articulatoires maximales du locuteur, lors de la réalisation de gestes vocaliques. Pour ce qui concerne des sujets ne souffrant d'aucune pathologie, de nombreuses études ([4] par exemple) ont montré qu'une augmentation de la vitesse d'élocution pouvait entraîner une compression des durées et une réduction de l'espace vocalique, c'est-à-dire une certaine centralisation des voyelles dans cet espace. Cependant, ce phénomène de centralisation n'a été observé que pour deux voyelles, en l'occurrence pour le [i] et pour le [u].

Qu'en sera-t-il pour des sujets bègues ou anciens bègues, sachant que la plupart des études [3] ont montré une centralisation du triangle vocalique déjà en vitesse d'élocution normale, pour les bègues (par rapport à des locuteurs de contrôle) ? De même, un travail plus récent de Blomgren *et al.* [1] confirme la réduction du triangle vocalique dans la parole fluente des bègues, bien qu'un travail de Prosek *et a*l. [5] infirme de telles conclusions, dans la mesure où leurs données ne présentent pas de centralisation vocalique, aussi bien en parole fluente qu'en parole disfluente.

L'intérêt de cette étude sera donc double : premièrement, il s'agira d'apporter des données supplémentaires afin de vérifier s'il y a ou non réduction du triangle vocalique dans la parole bègue ; deuxièmement de voir si un phénomène d' « undershoot » s'opère également pour ce type de locuteur en vitesse d'élocution rapide. Nos hypothèses majeures sont les suivantes : 1) on devrait observer un espace vocalique plus restreint en parole bégayée, par rapport à la parole non pathologique, étant donné la difficulté pour les bègues à gérer des faits coarticulatoires et en conséquence l'atteinte de « cibles » vocaliques adéquatement amples ; 2) il serait, en conséquence, difficile pour les locuteurs bègues de réduire davantage leur espace vocalique avec l'augmentation de la vitesse d'élocution, si celui-ci est déjà exigu.

2. Procédure expérimentale

2.1. Locuteurs et corpus

Neuf locuteurs adultes âgés de 25 à 30 ans, dont trois sans trouble de la parole, trois bègues et trois anciens bègues qui ont suivi une thérapie et qui ne présentent ni signe de bégayages ni stratégie d'évitement ont participé à cette étude. Tous avaient pour consigne de répéter à dix reprises des séquences [pVp] introduites dans des phrases porteuses, où [V] était soit [i], [a] ou [u]. Le corpus consistait donc à lire les phrases :

1. C'est une pipe ça.
2. C'est une pape ça
3. C'est une poupe ça.

Les répétitions ont été enregistrées avec un microphone Sennheiser e845 S relié à un portable PC (carte son RealTek AC97) et en utilisant Audacity (Fréquence d'échantillonage : 44100 Hz – 16 bits) comme logiciel d'acquisition.

2.2. Mesures acoustiques et calcul de l'espace vocalique

Mesures acoustiques

Les données recueillies ont été analysées à l'aide du logiciel Praat©. Seules les séquences fluentes ont été retenues pour cette étude. Les mesures de F1 et de F2 ont été prises au milieu de la structure formantique stable des voyelles.

Ces valeurs, qui représentent les résonances dans le conduit vocal, permettent de faire des inférences de la configuration du conduit vocal lors de la production d'une voyelle [6].

Calcul de l'espace vocalique

En plus de ces observations indirectes sur l'élévation et l'avancement de la langue dans la cavité buccale, l'aire du triangle vocalique sera également calculée [1]. Cette mesure, exprimée en Hz², permet d'obtenir une indication sur l'espace utilisé en vue de réaliser la distinction entre les voyelles. Il est à noter que l'aire des triangles vocaliques est obtenue par le calcul suivant :

Aire = $\sqrt{P(P-a)(P-b)(P-c)}$

où a, b et c représentent la distance entre les coordonnées de deux voyelles qui est quantifiée par :

$\sqrt{(xb - xa)^2 + (yb - ya)^2}$ (x étant la coordonnée de F1 d'une des trois voyelles et y la coordonnée de F2) et où P est le résultat de : (a + b + c) / 2

3. Résultats

Des analyses de variance (ANOVA) à 3 facteurs – *qualité vocalique* ([i, a, u]), *groupes* de locuteurs (bègues *vs*. locuteurs de contrôle *vs*. anciens bègues) et *vitesses* d'élocution (normale *vs*. rapide) - ont été effectuées sur les mesures de F1 et de F2. Pour la détermination de la signification statistique des *effets principaux*, ainsi que des *interactions*, seuls les résultats significatifs avec une probabilité de moins de 5% d'avoir été obtenus par chance ($p<0.05$) seront retenus. Les trois effets principaux n'ont pas révélé de significativité statistique globale. Cependant, des analyses ANOVA du *F2* ont été significatives pour le groupe de contrôle *vs*. les bègues et les anciens bègues *vs*. les bègues (quelles que soient la qualité vocalique et la vitesse d'élocution, $p<0.05$). Des ANOVA séparées sur la vitesse d'élocution indiquent un effet significatif pour les anciens bègues et le groupe de contrôle (quelle que soit la qualité vocalique, $p<0.05$).

3.1. Comparaison de la structure formantique en vitesse d'élocution normale

Structure formantique du [i]

La valeur moyenne de F1 a été mesurée à 218 Hz (E-T : 16 Hz) en moyenne, et celle de F2 à 2046 Hz (49 Hz) pour le locuteur de contrôle lorsqu'il répétait la voyelle [i]. Ce même son a été évalué avec une première zone de résonance à 273 Hz (7 Hz) et une deuxième à 1752 Hz (53) pour le locuteur bègue. Par conséquent, il est possible d'observer que la différence entre les deux productions provient surtout de F2, autrement dit de l'avancement de la langue dans la cavité buccale. Par inférence, il semblerait que le lieu d'articulation de la voyelle soit moins antérieur pour le locuteur bègue. En ce qui concerne le premier formant, les valeurs entre les deux locuteurs sont légèrement plus élevées pour B (218 Hz pour le locuteur de contrôle *vs*. 273 Hz pour le bègue), mais rappelons que l'effet principal était non significatif.

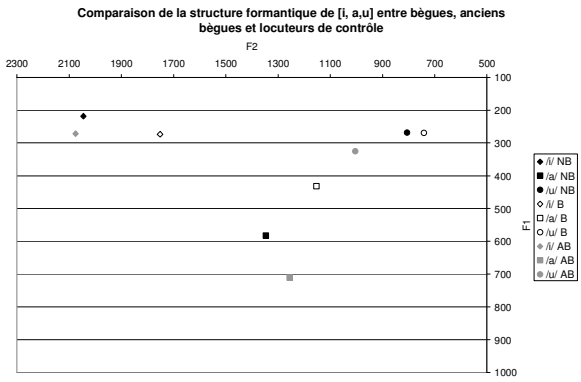


Figure 1 : comparaison des productions des voyelles [i, a, u] produites par un locuteur de contrôle (signes noirs), un locuteur bègue (signes vides) et un ancien bègue (signes gris).

Quant à l'étude de la structure formantique de la voyelle [i] chez l'ancien bègue, elle ne révèle pas de différences significatives avec les moyennes obtenues pour le locuteur de contrôle. En effet, l'opposition entre les valeurs F2 du bègue et du sujet non-pathologique ne semble pas pertinente, dans la mesure où le second formant a été quantifié à 2075 Hz (20 Hz) pour le sujet qui a suivi une thérapie (*vs*. 2046 Hz). Autrement dit, la position de la langue sur l'axe « antérieur – postérieur » semble identique pour le locuteur de contrôle et l'ancien bègue. Quant au premier formant, il est en moyenne de 272 Hz (9 Hz), c'est-à-dire qu'il est proche de ce qui a été constaté pour le bègue (273 Hz), et légèrement moins élevé, si on le compare à la moyenne obtenue pour le locuteur de contrôle (218 Hz).

Par conséquent, il est possible de conclure que la voyelle [i] est quasiment identique sur le plan qualitatif pour NB et pour AB et qu'elle est produite avec une position de la langue moins en avant pour B.

Structure formantique du [a]

Pour ce qui est du [a], réalisé par le sujet non-pathologique, on peut observer que le F1 est à 582 Hz (47 Hz) et le F2 à 1347 Hz (42 Hz). Ces valeurs ne sont pas celles prélevées pour le sujet bègue, tant pour le premier que pour le deuxième formant. En effet, si l'on prend le cas de F1, mesuré en moyenne à 432 Hz (26 Hz) pour le bègue, on peut noter une baisse sensible (même si l'effet principal n'avait pas été significatif) de cette valeur, ce qui traduirait le fait que la voyelle soit produite avec une aperture plus petite pour cette catégorie de sujet, par rapport au locuteur de contrôle. Il en est de même pour la valeur moyenne de F2, qui est de 1153 Hz (25 Hz) pour le locuteur bègue, alors que ce même paramètre était de 1347 Hz pour le sujet de contrôle. Quant aux valeurs obtenues pour l'ancien bègue, elles sont différentes à la fois du locuteur de contrôle et du sujet bègue. En effet, les résultats montrent que F1 est plus élevé pour ce groupe de locuteur, puisqu'il a été quantifié à 710 Hz (14 Hz), tandis qu'il a été mesuré à 432 Hz pour B et à 582 Hz pour NB. De même, la valeur F2 se trouve à un niveau intermédiaire par rapport aux deux autres locuteurs, étant donné qu'elle est de 1256 Hz (21 Hz) en moyenne (*vs*.

1347 Hz pour le locuteur NB et *vs*. 1153 Hz pour le locuteur B).

Structure formantique du [u]

Le premier formant de la voyelle [u] produite par le locuteur de contrôle a été observé à 268 Hz (23 Hz) en moyenne. Ce résultat est quasiment identique pour le sujet bègue, F1 ayant été mesuré à 269 Hz (11 Hz), et est moins élevé par rapport à l'ancien bègue, locuteur pour lequel la première zone de résonance était située à 325 Hz (8 Hz). C'est l'observation du deuxième formant qui révèle davantage de différences entre le locuteur de contrôle et l'ancien bègue. En effet, F2 a été mesuré à 805 Hz (50 Hz) pour le premier, alors qu'il est de 1004 Hz (32 Hz) pour le second. Il est important d'ajouter que la valeur de ce même paramètre chez le locuteur bègue est proche de celle de NB, étant donné qu'il a été mesuré à 741 Hz (50 Hz).

En conclusion, le [u] produit par le locuteur bègue est comparable à celui du sujet non-pathologique, si ce n'est qu'il est réalisé avec la langue placée légèrement plus en arrière. Quant aux réalisations de AB, elles suggèrent une antériorisation du [u] par rapport aux deux premiers locuteurs.

3.2. Effets de l'augmentation de la vitesse d'élocution selon le type de locuteurs

Conséquences de la vitesse d'élocution sur le [i]

L'augmentation de la vitesse d'élocution a pour effet d'accroître la valeur de F1 qui passe de 218 Hz à 253 Hz (14 Hz) lorsqu'il est demandé au locuteur de contrôle de parler plus rapidement. Parallèlement à cela, F2 diminue puisque la mesure pour la deuxième zone de résonance atteint 1989 Hz (37 Hz) en vitesse d'élocution rapide, alors que ce même paramètre était quantifié à 2046 Hz lorsque le sujet de contrôle devait parler à vitesse d'élocution normale. Il en va de même pour AB, étant donné que F1 augmente très légèrement lorsque ce dernier parle plus rapidement, en passant de 272 Hz à 291 Hz (25 Hz). Quant au deuxième formant du [i], il diminue également puisqu'il était de 2075 Hz en vitesse d'élocution normale et qu'il se trouve à 2032 Hz (27 Hz) lorsque ce locuteur parle plus rapidement. Cette tendance ne se confirme pas pour le locuteur bègue. En effet, il est possible de constater une stabilité des valeurs de F1 (273 Hz en vitesse d'élocution normale *vs*. 263 Hz (14 Hz) en vitesse d'élocution rapide) et de F2 (1752 Hz *vs*. 1772 Hz (205 Hz)). Par conséquent, il n'y a pas de différence significative entre les productions de la voyelle [i] en vitesses d'élocution normale et rapide, chez les bègues.

Conséquences de la vitesse d'élocution sur le [a]

Nous avons pu observer plus haut que le F1 de la voyelle produite par le locuteur de contrôle était de 582 Hz et que le F2 était évalué à 1347 Hz. Lorsque la vitesse d'élocution augmente, c'est principalement la valeur du premier formant qui se trouve modifiée, étant donné qu'il passe de 582 Hz à 517 Hz (27 Hz). Quant à la deuxième zone de résonance, elle reste stable (1347 Hz en vitesse d'élocution normale *vs*. 1296 Hz (43 Hz)). Peu d'évolutions ont été observées pour la même voyelle chez l'ancien bègue. Ainsi, la valeur de F1 reste stable en passant de 710 Hz à 681 Hz (29 Hz) lorsque le locuteur a pour consigne d'accélérer la vitesse d'élocution, tout comme la moyenne pour F2, puisqu'elle est de 1256 Hz en vitesse d'élocution normale et de 1297 Hz (33 Hz) en vitesse d'élocution rapide. De même, aucune différence pertinente n'a été constatée lorsque la vitesse d'élocution du locuteur bègue augmente : F1 passe de 432 Hz à 438 Hz (35 Hz) et F2 de 1153 Hz à 1116 Hz (20 Hz). Il est à noter que ces résultats sont conformes à la littérature, étant donné que Lindblom [4] et Ferbach-Hecker [2] n'ont pas observé de centralisation pour le [a].

Conséquences de la vitesse d'élocution sur le [u]

L'accélération de la vitesse d'élocution entraine chez NB une légère modification de la structure formantique du [u]. En effet, on peut constater que F2 passe de 805 Hz à 886 Hz (97 Hz). Cependant, cette différence est non significative (à cause d'un écart-type relativement trop élevé) et doit, en conséquence, être prise avec précaution. Quant à la première zone de résonance, elle est située à 268 Hz, lorsqu'aucune contrainte de vitesse n'est exigée et elle est de 292 Hz (11 Hz) lorsque la vitesse d'élocution est accélérée. Des remarques similaires peuvent être effectuées pour AB, étant donné que F2 était de 1004 Hz en vitesse d'élocution normale et qu'il est évalué à 1063 Hz (75 Hz) en vitesse d'élocution rapide.

Enfin, il est intéressant de noter que les valeurs F1 et F2 du [u] sont quasiment identiques pour le locuteur bègue dans les deux vitesses d'élocution, puisque F1 est de 269 Hz en vitesse d'élocution normale et de 258 Hz (29 Hz) en parole rapide et que F2 est à 741 Hz lorsque le locuteur parle sans consignes temporelles et de 724 Hz (66 Hz) lorsqu'il lui est demandé de parler vite.

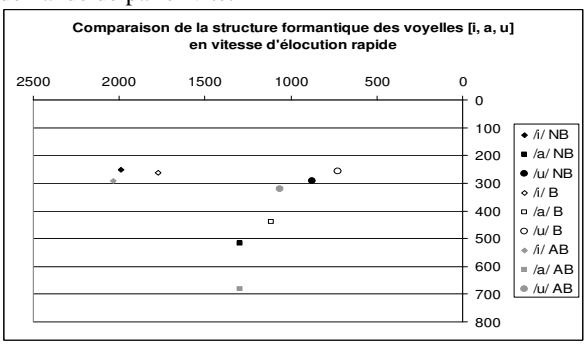


Figure 2 : comparaison des productions des voyelles [i, a, u] produites par un locuteur de contrôle (signes noirs), un locuteur bègue (signes vides) et un ancien bègue (signes gris) en vitesse d'élocution rapide.

3.3. Comparaison de l'aire des triangles vocaliques

Il est important de signaler que le calcul de l'aire et la comparaison des triangles vocaliques n'a pas de signification fonctionnelle à proprement parler. Elle permet cependant d'obtenir une vue globale sur l'espace vocalique maximal utilisé pour obtenir une opposition qualitative entre les voyelles et sur l'amplitude des mouvements réalisés à

cette fin. Les valeurs obtenues pour les trois locuteurs qui font l'objet de cette étude sont données dans le tableau 1.

Tableau 1 : Comparaison de l'aire des triangles vocaliques entre le locuteur de contrôle (NB), le sujet bègue (B) et l'ancien bègue (AB).

Aire (Hz²)	NB	B	AB
Vitesse d'élocution normale	208.072	81.722	212.856
Vitesse d'élocution rapide	164.620	93.112	173.343

L'étude de l'aire des triangles vocaliques révèle que la différenciation des voyelles s'effectue, chez les bègues, sur une surface deux fois moins élevée par rapport au locuteur de contrôle, puisque l'espace vocalique a été calculé à 81.722 Hz² pour le premier, alors qu'il est de 208.072 Hz² pour le second. Par ailleurs, il est également possible de constater que l'aire obtenue pour les voyelles réalisées par AB est comparable au sujet non-pathologique, étant donné qu'elle est de 212.856 Hz² (*vs*. 208.072 Hz² pour NB).

En vitesse d'élocution rapide, il est possible de constater une réduction significative de l'aire du triangle vocalique pour NB, puisque cette valeur passe de 208.072 Hz² à 164.620 H². Ce constat vaut également pour AB, la surface du triangle étant de 212.856 Hz² en vitesse d'élocution normale et de 173.343 Hz² lorsque le locuteur parle plus rapidement.

Pour le locuteur bègue cependant, aucune réduction de l'aire du triangle vocalique n'est constatée. Au contraire, celle-ci augmente très légèrement, en passant de 81.722 Hz² à 93.112 Hz².

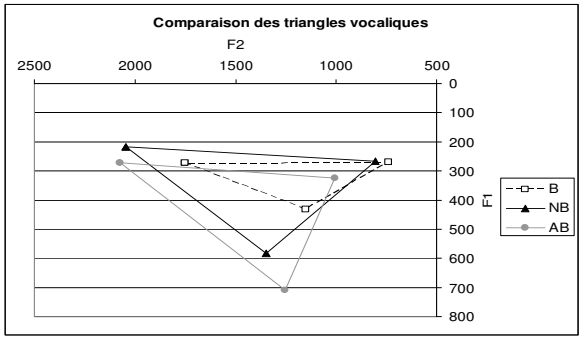


Figure 3 : Comparaison de l'aire des triangles vocaliques entre le locuteur de contrôle (NB), le sujet bègue (B) et l'ancien bègue (AB) en vitesse d'élocution normale.

4. Synthèse et conclusion

La structure formantique des voyelles [i, a, u] produites par l'ancien bègue est comparable à celle obtenue pour le locuteur de contrôle. Il semble alors cohérent de constater que l'aire des triangles vocalique de ces deux locuteurs soit similaire.

Ce n'est cependant pas le cas pour le locuteur bègue, pour qui l'aire du triangle est fortement restreinte, montrant ainsi une différence de stratégie par rapport aux locuteurs de contrôle et anciennement bègues ; c'est le F2 qui est le grand responsable des différences observées dans les espaces vocaliques. Ce résultat suggère une particularité dans la gestion du lieu d'articulation chez les bègues [1]. Il est également intéressant de noter que l'augmentation de la vitesse d'élocution entraîne, pour les contrôles et les anciens bègues, une diminution de l'aire du triangle. Autrement dit, le locuteur de contrôle et l'ancien bègue utilisent un phénomène d' « undershoot » en accélérant leur vitesse d'élocution, ce qui n'est pas le cas pour le locuteur bègue, étant donné que la structure formantique des voyelles reste stable malgré le changement de rythme. Par conséquent, on peut dire que le locuteur bègue « sait » moins utiliser les possibilités de variation de l'espace vocalique, suivant le contexte de vitesse d'élocution.

Remerciements

Nous remercions Rudolph Sock pour ses remarques et ses suggestions. Cette recherche a été financée en partie par le Programme de Recherche ACI TTT 2003-2006 du Ministère de la Recherche et des Nouvelles Technologies, ainsi que par le Programme MISHA attribué à la Composante Parole et Cognition de l'E.A. 1339 – LiLPa.

Bibliographie

[1] M. Blomgren M. Robb and Y. Chen. A note on Vowel Centralization in Stuttering and Nonstuttering Individuals. In *Journal of Speech, Language, and Hearing Research*, volume 41, pages 1042-1051, 1998.

[2] V. Ferbach-Hecker. La résistivité de la qualité des voyelles orales du français. In *SCOLIA*, volume 20, pages 115-134, 2005.

[3] R. Klich and G. May Spectrographic study of vowels in stutterers' fluent speech. In Journal of Speech and Hearing Research, volume 25, pages 364-370, 1982.

[4] B. Lindblom. On vowel reduction. In *The Royal Institute of Technology, Speech Transmission Laboratory*, volume 29, 1963.

[5] R. Prosek A. Montgomery B. Walden and D. Hawkins. Formant frequenciezs of stuttered and fluent vowels. Journal of Speech and Hearing Research, volume 30, pages 301-305, 1987.

[6] K. Stevens and A. House. Development of a description of vowel articulation. In Journal of Acoustical Society of America, volume 27, pages 484-493, 1955.

[7] B. Vaxelaire. Etude comparée des effets des variations de debit –lent, rapide- sur les paramètres articulatoires, à partir de la cinéradiographie (sujets français). Thèse de Doctorat Nouveau Régime, Université des Sciences Humaines de Strasbourg, 1993.

Modélisation Statistique et Informations Pertinentes pour la Caractérisation des Voix Pathologiques (Dysphonies)

G. Pouchoulin[1], *C. Fredouille*[1], *J.-F. Bonastre*[1], *A. Ghio*[2], *M. Azzarello*[3], *A. Giovanni*[3]

[1]LIA, Avignon (France)
[2]LPL-CNRS, Aix en Provence (France)
[3]LAPEC, Marseille (France)

ABSTRACT

This paper investigates the class of information relevant for the task of automatic classification of pathological voices. By using a GMM-based classification system (derived from the Automatic Speaker Recognition domain), the focus was made on three main classes of information : energetic, voiced, and phonetic information. Experiments made on a pathological corpus (dysphonia) have shown that phonetic information is particularly interesting in this context since it permits to refine the selection of the relevant information by looking at phonem- or phonem class-level (e.g. nasal vowels).

1. INTRODUCTION

Dans le domaine de la phoniatrie, l'évaluation de la qualité vocale est un sujet sensible, au centre de nombreuses études dans des domaines multi-disciplinaires [11][7]. Dans le cas de dysphonies (altération du son laryngé), sur lesquelles se concentre cet article, le dysfonctionnement vocal peut être évalué suivant deux approches, que sont l'analyse perceptive et l'analyse objective.
L'analyse perceptive ou auditive est la plus utilisée en pratique clinique. Elle consiste à caractériser la qualité vocale par une simple écoute attentive ; de par sa subjectivité, le recours à des jury d'experts peut être nécessaire afin d'augmenter la fiabilité de l'analyse. Les inconvénients majeurs de cette approche sont le manque de fiabilité dû à différents facteurs de variabilité, qu'elle soit intra ou inter-individuelle, ainsi que la difficulté de mise en place dans le cas d'un jury d'écoute (réunions périodiques de plusieurs experts, durée des séances d'écoute, ...).
L'analyse objective ou instrumentale (comme le système EVA^{TM} [10], Evaluation Vocale Assistée - SQLab) s'appuie sur l'acquisition de nombreuses données quantitatives (mesures acoustiques, aérodynamiques et physiologiques) par le biais d'appareils de mesure. Elle offre une approche complémentaire à l'examen laryngoscopique et à l'interrogatoire du patient effectués par les praticiens. Dans [12], 86% de concordance entre l'analyse perceptive et objective sont atteints en utilisant 9 paramètres acoustiques et aérodynamiques (F0, intensité, jitter, coefficient de Lyapunov, rapport signal sur bruit, débit d'air buccal, pression sous-glottique estimée, étendue vocale, temps maximum de phonation). L'approche instrumentale offre donc des résultats très acceptables mais encore insuffisants pour son usage dans une pratique clinique quotidienne. Par ailleurs, l'acquisition des mesures sur le /a/ tenu (généralement utilisée dans ce cas) reste controversée dans la littérature [8] car elle tend à sous-estimer la dysphonie. Des mesures effectuées sur de la parole continue permettraient de prendre en compte par exemple, les phénomènes vocaux de l'attaque, reconnus comme pertinents dans l'évaluation des dysphonies.
En résumé, à ce jour, aucune approche d'évaluation de la qualité de la voix ne semble répondre correctement à l'attente des cliniciens, même si l'analyse perceptive reste incontournable dans la pratique quotidienne. En outre, cette dernière reste la seule référence à laquelle sont confrontées les méthodes objectives. Dans la continuité d'une étude préliminaire [4] menée par le LIA en vue d'adapter des techniques de reconnaissance automatique du locuteur (RAL) à l'évaluation de voix pathologiques (dysphonies), notre objectif est de proposer une méthode instrumentale mieux adaptée au suivi de la pathologie des patients : facilité et rapidité d'utilisation, non contraignante pour le patient et accessible pour les cliniciens. Comparée à des méthodes instrumentales classiques, les originalités de cette approche basée sur une modélisation statistique, reposent sur :

- sa capacité à analyser de la parole continue (et non des voyelles tenues) proche de l'élocution naturelle ;
- sa capacité à traiter de grands corpus, permettant de mener des études à grandes échelles et d'obtenir des informations statistiques significatives ;
- une analyse acoustique, simple et automatique, permettant une simplicité d'instrumentation et un faible coût humain.

Le système conçu pour cette tâche particulière s'appuie sur l'approche à base de GMM, état de l'art pour la RAL. Il est issu des outils de RAL, disponibles en « version libre » (LIA_SpkDet et ALIZE) et développés au LIA.
Dans la continuité de ce travail, nous proposons ici une première étude sur l'extraction des informations utiles à la caractérisation des voix pathologiques. Nous nous intéresserons plus particulièrement à trois sources d'informations : (1) signal complet nettoyé des parties silencieuses, (2) segments phonétiques, (3) segments voisés uniquement.
Le corpus dysphonique utilisé dans cette étude ainsi que le système de classification automatique sont décrits en sections 2 et 3. La sélection des différentes classes d'information est détaillée en section 4, puis évaluée dans un contexte expérimental en section 5. Finalement, la section 6 fournit une conclusion à ce travail.

2. EVALUATION DES DYSPHONIES

Dans cette étude, nous nous intéressons à des dysphonies d'origine multiple (nodules, polypes, oedèmes, kystes...) classées perceptivement selon le paramètre G de l'échelle

d'évaluation GRBAS de Hirano [5]. Le corpus mis à disposition par le LAPEC (Hôpital de la Timone Marseille) est constitué de 80 échantillons de voix de femmes correspondant à 20 sujets témoins et 60 patientes dysphoniques, âgées de 17 à 50 ans (moyenne de 32.2 ans). Chaque sujet a été enregistré sur la lecture d'un paragraphe de « La chèvre de Monsieur Seguin » d'Alphonse Daudet. Les enregistrements ont été évalués selon le grade global G de dysphonie de l'échelle GRBAS par un jury d'experts ; les décisions ont été prises par consensus afin d'en limiter la variabilité inter-auditeur et en une seule séance afin d'en limiter la variabilité intra-auditeur. L'ensemble du corpus étiqueté se présente donc de la manière suivante : 20 voix normales G0, 20 voix présentant une dysphonie légère G1, 20 voix une dysphonie moyenne G2 et 20 voix une dysphonie sévère G3.

3. SYSTÈME DE CLASSIFICATION DES VOIX PATHOLOGIQUES

Le principe retenu consiste en l'adaptation d'un système état de l'art de RAL à la classification de voix pathologiques suivant leur degré de dysphonie. Trois phases sont nécessaires et sont décrites dans les sections suivantes.

3.1. Extraction de l'information acoustique

L'extraction de l'information acoustique est issue des méthodes courantes employées en RAL : pour chaque trame de signal analysé (fenêtre de Hamming de 20ms avec un pas de 10ms), sont extraits 16 coefficients cepstraux (MFCC) obtenus à partir de 24 coefficients de banc de filtres répartis sur une échelle de MEL. Les dérivées premières des coefficients MFCC (Δ) sont ajoutées aux vecteurs de paramètres qui sont finalement normalisés pour obtenir une distribution de moyenne 0 et variance 1 (les moyennes et les variances sont estimées sur les portions jugées utiles du signal dont la sélection est décrite en section 4).

3.2. Modélisation

En RAL, l'état de l'art repose sur une modélisation statistique (GMM : Gaussian Mixture Model)[1]. Un GMM *X* est une somme pondérée de M distributions gaussiennes multidimensionnelles, chacune caractérisée par un vecteur moyen $\overline{x}$ de dimension *d*, une matrice de covariance Σ de dimension *d* x *d* et un poids p de la gaussienne au sein de la mixture. Durant la phase d'apprentissage, les paramètres ($\overline{x}$, Σ, p) sont estimés par l'algorithme EM/ML[1].
Classiquement, deux phases d'apprentissage sont nécessaires en RAL pour pallier le manque de données d'apprentissage disponibles pour chaque locuteur [1] :

. apprentissage d'un modèle générique (aussi appelé « modèle du monde ») estimé par l'algorithme EM/ML sur une grande quantité de données (population de locuteurs) ;
. apprentissage du modèle locuteur dérivé du modèle du monde par application des techniques d'adaptation (MAP, Maximum a Posteriori) [9].

Dans le contexte pathologique, un modèle ne correspond plus à un locuteur donné mais à un niveau de sévérité de dysphonie. Il sera appelé **modèle de grade** G_g avec $g \in \{0, 1, 2, 3\}$. Le modèle de grade est appris en utilisant l'ensemble des locuteurs de même grade. On s'assurera que les voix utilisées pour l'apprentissage des modèles de grade, sont exclues des jeux de tests afin de différencier la détection de la pathologie de la reconnaissance du locuteur (mise en oeuvre de la technique leave_x_out).
Les modèles GMM représentant les grades pathologiques sont établis comme suit :

. un modèle GMM générique est d'abord estimé par l'algorithme EM/ML sur un corpus français composé de 76 enregistrements de 2 mn chacun de voix de femmes exclusivement ; Cette population est extraite du corpus BREF [6] entièrement disjoint du corpus dysphonique
. les modèles de grade sont ensuite dérivés du modèle GMM générique suivant la technique d'adaptation MAP [9]. Seules les moyennes sont adaptées.

Tous les modèles GMM se composent de 128 composantes gaussiennes à matrices de covariance diagonales.

3.3. Classification

Lors de la phase de test, une mesure de similarité entre des vecteurs acoustiques y_t issus d'un signal et un modèle *X* est calculée suivant : $L(y_t|X) = \sum_{i=1}^{M} p_i \, L_i(y_t)$
où $L_i(y_t)$ est la vraisemblance du signal y_t par rapport à la gaussienne i, M le nombre de gaussiennes et p_i le poids de la gaussienne.
La **décision** correspondra au grade *g* du modèle G_g sur lequel la plus grande vraisemblance est obtenue. Cette définition de la décision est proche de celle de la tâche d'identification automatique du locuteur. On dira que le système a classé la voix du locuteur *Y* dans le grade *g*.

4. SÉLECTION DES SEGMENTS PERTINENTS

Dans ce papier, nous nous intéressons aux informations pertinentes pour la caractérisation des dysphonies. Trois niveaux d'extraction de l'information utile sont étudiés :

. « segments énergétiques » : le signal de parole est nettoyé des trames de silence (système de détection « parole/non parole » du LIA basé sur une modélisation statistique de l'énergie) ;
. « segments phonétiques » : extraite d'un alignement phonétique automatique contraint par le texte (système d'alignement du LIA basé sur un décodage Viterbi, à partir d'un lexique de mots et leurs variantes phonologiques - 38 phonèmes du français) ;
. « segments voisés » : extraction des sons de parole voisés par analyse de la fréquence fondamentale F0 (logiciel PRAAT [2] sur l'intervalle de fréquence [75,600] Hz).

Selon le niveau d'analyse choisi, est opérée une sélection de trames issues du signal qui sera utilisée lors la normalisation des paramétres acoustiques, l'apprentissage des modèles et la décision.

5. EXPÉRIENCES

Les expériences ont été réalisées après « adaptation » du système de RAL du LIA. Ce système (appelé LIA_SpkDet) repose entièrement sur la plateforme libre ALIZE [3] conçue et réalisée dans le cadre du programme Technolangue. LIA_SpkDet est également distribué en logiciel libre.

[1]Expectation-Maximization/Maximum Likelihood

	Grade 0	Grade 1	Grade 2	Grade 3	Total
Information	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/80)
Energétique	95,0 (19)	70,0 (14)	50,0 (10)	60,0 (12)	68,75 (55)
Voisement	95,0 (19)	65,0 (13)	50,0 (10)	75,0 (15)	71,25 (57)
Phonétique	95,0 (19)	60,0 (12)	55,0 (11)	75,0 (15)	71,25 (57)

TAB. 1: Résultats de la classification 4-Grades suivant différentes classes d'informations extraites

Classification	Grade 0	Grade 1	Grade 2	Grade 3
Locuteurs de Grade 0	**19**	1	0	0
Locuteurs de Grade 1	2	**12**	4	2
Locuteurs de Grade 2	2	5	**11**	2
Locuteurs de Grade 3	0	1	4	**15**

TAB. 2: Matrice de confusion - Phonétique

5.1. Protocole expérimental

Il s'agit de classer une voix suivant les 4 niveaux du grade global de l'échelle GRBAS. Par conséquent, 4 modèles de grade G_g sont à estimer avec $g \in \{0, 1, 2, 3\}$.
Lors de la phase de test, la mise en oeuvre de la technique leave_x_out (en vue de séparer les données de test et d'apprentissage) permet de comparer chaque voix y_t de grade g avec :
- 1 modèle G_g appris à partir des 19 voix restantes de grade g ($y_t \notin$ aux 19 voix) ;
- 3 x 20 modèles $G_{\overline{g}}$ appris chacun à partir de 19 voix de grade $\overline{g}$ avec $\overline{g} \in \{0, 1, 2, 3\} - \{g\}$.

A l'issu de ces comparaisons, les moyennes des vraisemblances des tests sur chaque grade (1 modèle G_g et 3 x 20 modèles $G_{\overline{g}}$) sont calculées et comparées pour fournir une unique décision pour la voix (y_t) de grade g.
Note : Le même nombre de voix (19) est utilisé pour l'ensemble des modèles de grade.

5.2. Résultats

Classes d'informations utiles
Le tableau 1 donne les résultats des différentes classes d'informations sélectionnées en vue de la caractérisation des voix pathologiques. Les expériences relatives aux segments « voisés » et « phonétiques » obtiennent le meilleur résultat (71,25% de réussite). Quelle que soit la classe d'information, on peut remarquer que le grade 0 est le mieux reconnu (95,0%) et que la confusion provient principalement des grades 1 et 2 (voir matrices de confusion Tab. 2 et 3). Concernant les segments « énergétiques », à durée équivalente avec les segments « phonétiques », la classification en grade 3 est dégradée, pouvant montrer une faiblesse du détecteur parole/non parole sur des dysphonies très sévères. Les informations « voisées » obtiennent un taux de réussite équivalent aux informations « phonétiques » alors que leurs durées sont plus courtes de 10 à 20% suivant les grades. Cette observation tend à démontrer que les informations non voisées présentes dans la classe des informations « phonétiques » sont moins pertinentes, ce qui semble logique dans le contexte d'une analyse des dysphonies qui concernent une altération du voisement.
Note : Tous les résultats fournis dans ce papier sont issus du classifieur GMM et doivent être interprétés d'un point de vue statistique.

Classification	Grade 0	Grade 1	Grade 2	Grade 3
Locuteurs de Grade 0	**19**	1	0	0
Locuteurs de Grade 1	3	**13**	2	2
Locuteurs de Grade 2	3	5	**10**	2
Locuteurs de Grade 3	0	1	4	**15**

TAB. 3: Matrice de confusion - Voisement

Analyse phonétique
Dans [7], une étude descriptive et perceptive des caractéristiques pathologiques de chaque phonème constitutif d'un échantillon de parole chez des patients dysphoniques est proposée sous le nom de "phonetic labelling". Nous proposons ici de suivre une démarche similaire en observant le comportement du système de classification automatique suivant différentes classes de phonèmes.
A partir des résultats obtenus sur les segments « phonétiques », une analyse a été réalisée sur la pertinence des différentes classes de phonèmes dans la classification des voix pathologiques. Il est à noter que cette analyse porte uniquement sur le pouvoir décisionnel de ces différentes classes (cette catégorisation n'est pas utilisée lors des phases de normalisation des paramètres ni d'apprentissage pour lesquelles l'ensemble des informations « phonétiques » a été utilisé). Le tableau 4 présente une première analyse des décisions suivant une catégorisation détaillée « Consonnes/Voyelles ». Il est intéressant de remarquer :
. la pertinence des différentes classes pour le grade 0 (85% en moyenne de bonne classification) ;
. des différences marquées entre les classes pour le grade 3 (de 75 à 85% pour les voyelles orales, les semi-consonnes et les occlusives contre 45 à 60% pour les voyelles nasales, les consonnes nasales, liquides et fricatives) ;
. pour les grades 1 et 2, les consonnes occlusives et les voyelles nasales apportent peu d'information (de 35% à 45% uniquement de bonne classification). En revanche, les semi-consonnes, liquides et fricatives réagissent plus favorablement au grade 2 (de 60% à 65% de bonne classification) ; les voyelles orales et consonnes nasales plus favorablement au grade 1 (60%) ;
. les semi-consonnes, malgré leurs courtes durées (de 0.39s à 0.44s), obtiennent un taux global de réussite de 67.5% contre 71.25% sur la totalité des informations « phonétiques ».

Un premier approfondissement de cette étude montre qu'au sein d'une même classe de phonèmes, les comportements peuvent être très différents suivant les grades. Par exemple, le tableau 5 présente les résultats de classification des voyelles nasales.
Note : Dans ce cas, la décision est prise au niveau du phonème uniquement.
On peut observer, par exemple, pour le grade 3 des taux de réussite variant de 25% pour le phonème [$\tilde{\varepsilon}$] à 70% pour le phonème [$\tilde{\text{œ}}$]. Ce même comportement a été observé sur d'autres classes de phonèmes.
Par ailleurs, des travaux en cours devraient permettre d'établir des corrélations entre ces résultats et ceux obtenus par la méthode de « phonetic labelling »(analyse perceptive).

6. DISCUSSION

Dans ce papier, nous proposons une analyse de l'information pertinente, véhiculée par le signal de parole, pour la caractérisation des voix pathologiques. Trois niveaux

	Grade 0	Grade 1	Grade 2	Grade 3	Total
Classe phonét.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/80)
Voyelle	**95,0 (19)**	**60,0 (12)**	**35,0 (7)**	**70,0 (14)**	**65,00 (52)**
Voyelle orale	95,0 (19) 4,14 (s)	60,0 (12) 4,09 (s)	45,0 (9) 4,15 (s)	75,0 (15) 4,12 (s)	68,75 (55)
Voyelle nasale	95,0 (19) 0,89 (s)	40,0 (8) 0,85 (s)	35,0 (7) 0,85 (s)	45,0 (9) 0,66 (s)	53,75 (43)
Consonne	90,0 (18)	50,0 (10)	60,0 (12)	65,0 (13)	66,25 (53)
Semi -consonne	90,0 (18) 0,39 (s)	35,0 (7) 0,43 (s)	60,0 (12) 0,43 (s)	85,0 (17) 0,44 (s)	67,50 (54)
Consonne liquide	80,0 (16) 1,68 (s)	45,0 (9) 1,66 (s)	60,0 (12) 1,72 (s)	60,0 (12) 1,86 (s)	61,25 (49)
Consonne nasale	75,0 (15) 1,40 (s)	60,0 (12) 1,41 (s)	50,0 (10) 1,57 (s)	50,0 (10) 1,46 (s)	58,75 (47)
Consonne fricative	90,0 (18) 1,53 (s)	40,0 (8) 1,56 (s)	65,0 (13) 1,64 (s)	45,0 (9) 1,77 (s)	60,00 (48)
Consonne occlusive	85,0 (17) 2,01 (s)	45,0 (9) 2,06 (s)	45,0 (9) 2,18 (s)	85,0 (17) 2,25 (s)	65,00 (52)

TAB. 4: Analyse par classe phonétique : % de réussite et durée moyenne par classe et par grade

Phonèmes	Grade 0	Grade 1	Grade 2	Grade 3
[ɑ̃]	**68,3 (41/60)**	**28,3 (17/60)**	**41,7 (25/60)**	**36,7 (22/60)**
[ɛ̃]	**47,5 (19/40)**	**35,0 (14/40)**	**35,0 (14/40)**	**25,0 (10/40)**
[ɔ̃]	**85,0 (51/60)**	**28,3 (17/60)**	**28,3 (17/60)**	**51,7 (31/60)**
[œ̃]	**75,0 (15/20)**	**30,0 (6/20)**	**25,0 (5/20)**	**70,0 (14/20)**

TAB. 5: Résultats de la classification 4-Grades suivant les voyelles nasales (décision au niveau du phonème)

d'informations ont été testés : segments « énergétiques » , « phonétiques » et « voisés » . Cette étude a été menée en utilisant un système de classification de voix pathologiques dérivé du domaine de la RAL et basé sur une approche statistique (GMM).

D'un point de vue expérimental, les informations « phonétiques » semblent les plus intéressantes dans cette étude, au sens où elles permettent d'affiner la sélection de l'information utile. En effet, l'étude des différentes classes phonétiques a montré l'influence de certains phonèmes ou classes de phonèmes dans la tâche de classification des voix pathologiques (68.75% de réussite pour les voyelles orales contre 53.75% pour les voyelles nasales). Néanmoins, nous avons montré également que des comportements différents peuvent intervenir au sein d'une même classe. Par ailleurs, il est intéressant de constater que certaines classes de phonèmes sont plus discriminantes que d'autres pour les grades 1 et 2, qui restent problématiques dans le cadre d'une décision globale (totalité de l'information présente). Il est à noter cependant que la petite taille du corpus (80 voix dont 60 dysphoniques) est à prendre en compte dans la validité de ces résultats ainsi que les caractéristiques intrinsèques du système de classification utilisé.

Il serait à présent intéressant de définir un paradigme de décision basé sur les informations phonétiques, permettant d'améliorer les performances du système automatique. La définition d'un arbre de décision phonétique constitue une voie intéressante pour améliorer la fiabilité de la classification.

7. CONCLUSION

Si d'autres types d'analyses instrumentales permettent de meilleurs résultats pour le moment [12], le taux de classification correcte de 70% avec 80 locuteurs et 4 classes est encourageant. En effet, les performances du système peuvent être améliorées dans plusieurs directions : (1) l'augmentation du corpus d'apprentissage (élément très important dans les systèmes de RAL), (2) une extraction d'information acoustique mieux adaptée à l'analyse des dysphonies, (3) une adaptation sélective des segments pertinents. Enfin, nous avons conscience que l'intérêt majeur de ce type d'outil de classification automatique est un certain déterminisme qui fait actuellement défaut à l'analyse perceptive. Cet outil restera un instrument d'évaluation et non un outil de décision qui reste clairement entre les mains du clinicien.

REMERCIEMENTS

Les auteurs tiennent à remercier le Laboratoire Audio-Phonologie Expérimentale et Clinique (LAPEC - Hôpital La Timone-Marseille) d'avoir mis à leur disposition le corpus dysphonique utilisé dans cette étude.

RÉFÉRENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds. A tutorial on text-independent speaker verification. In *EURASIP Journal on Applied Signal Processing*, volume 39, pages 430–451, 2004.

[2] P. Boersma and D. Weenink. Praat : doing phonetics by computer. http ://www.praat.org/.

[3] J.-F. Bonastre, F. Wils, and S. Meignier. Alize, a free toolkit for speaker recognition. In *ICASSP-05, Philadelphia, USA*, volume 39, pages 430–451, 2005.

[4] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio. Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia). In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 05)*, 2005.

[5] M. Hirano. Psycho-acoustic evaluation of voice : Grbas scale for evaluating the hoarse voice. In *Clinical Examination of voice, Springer Verlag*, 1981.

[6] L. Lamel, J. Gauvain, and L. Eskénazi. Bref, a large vocabulary spoken corpus for french. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99)*, 1991.

[7] J. Revis. L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale. In *Phd thesis, Université de la Méditerranée*, 2004.

[8] J. Revis, A. Giovanni, FL. Wuyts, and J.M. Triglia. Comparaison of different voice samples for perceptual analysis. In *Folia Phoniatr Logop*, pages 108–116, 1999.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models, digital signal processing (dsp). In *a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3)*, pages 19–41, 2000.

[10] B. Teston and B. Galindo. A diagnosis of rehabilitation aid workstation for speech and voice pathologies. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 95)*, pages 1883–1886, 1995.

[11] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning. The dysphonia severity index : an objective measure of vocal quality based on a multiparameter approach. In *Journal of Speech, Language, and Hearing Research 43*, pages 796–809, 2000.

[12] P. Yu, M. Ouakine, J. Revis, and A. Giovanni. Objective voice analysis for dysphonic patients : a multiparametric protocol including acoustic and aerodynamic measurements. In *Journal Voice 15*, pages 529–542, 2001.

Parler femme et *parler homme* en japonais actuel: Formes terminales et indices prosodiques

Yukihiro NISHINUMA, Akiko HAYASHI** & Hiroko YABE****

* CNRS, Laboratoire Parole et Langage, Aix-en-Provence, France
** Faculte des lettres, Universite Chuo, Tokyo, Japon
*** Tokyo Gakugei Universite, Tokyo, Japon
yukihiro.nishinuma@lpl.univ-aix.fr.fr

ABSTRACT

This work reports findings on the relationship between speaker sex and linguistic behavior of young Japanese in explanation-giving dialogues. The relationship between speaker sex and (a) the choice of utterance final forms; (b) the prosodic characteristics on these forms, has thus been examined. Data from 110 students of the Tokyo area revealed no statistically significant effect of the sex factor in the linguistic forms used. However utterance final syllables had a statistically significant effect both on intonation and on rhythm.

1. INTRODUCTION

Il existe des langues dans lesquelles les différences spécifiques au sexe sont plus ou moins codifiées. Le japonais en fait partie, et l'intègre dans son fonctionnement linguistique. Toutefois cette codification subissant une certaine mutation, ne correspond plus à l'état actuel de cette langue.

En effet, depuis les années 80, la recherche linguistique textuelle s'est orientée vers l'analyse du corpus des discours naturels, ce qui a permis d'éclaircir la réalité déviée de la norme des particules finales d'énoncé entre le «parler masculin» vs le «parler-féminin» [1]. En analysant le langage sur le lieu du travail, Ozaki a montré que la particule finale "~wa" n'était plus monopolisée par les femmes, et que la particule "~dawa" disparaissait, ou devenait marqueur de la vieille génération féminine [2]. Les particules finales "~da" et "~nda" entrent dans le parler des femmes et même cette tendance n'a plus de marque de sexe ; en revanche, la particule finale "~no", bien que très peu utilisée par les jeunes filles en conversation informelle reste une des formes féminines [3].

La particule finale "~wa" chez les femmes se réalise normalement en intonation montante Mcgloin [4]. Selon notre observation, la jeunesse actuelle n'hésitant plus à employer "~wa" en intonation descendante, il semble qu'il en soit fait une utilisation partagée selon les circonstances. D'autre part, Suzuki qui a examiné l'intonation des particules finales "~yo, ~no, ~noyo, ~nda/~noda" dans les émissions télévisuelles, les a classées en cinq catégories intonatives [5]. Malheureusement aucune donnée de fréquence fondamentale n'est présentée.

Comme nous venons de le voir, la neutralisation des spécificités dues au sexe sur les particules finales d'énoncé s'accentue dans la jeunesse japonaise actuelle, et il existe des variations prosodiques sur ces éléments syntaxiques. Nous nous posons alors les deux questions suivantes :
(1) Y a-t-il une préférence sélective des particules finales d'énoncé entre les sujets féminins et masculins ?
(2) Y a-t-il des différences prosodiques entre les sujets féminins et masculins sur les mêmes particules finales choisies ?

Dans la question 1, on examine d'abord le degré de neutralisation des différences suivant le sexe, et dans la seconde, on vérifie s'il n'existerait pas éventuellement, comme phénomène de compensation, des traces de dichotomie du sexe en réalisation phonétique, plus particulièrement en prosodie.

2. PROCEDURE EXPERIMENTALE

2.1. *Mode écrit : Composition des dialogues*

Afin d'obtenir des éléments de réponse à ces questions, deux expériences ont été mises en chantier. La première, linguistique, consistait à faire concevoir de courts dialogues comportant les particules finale "~nda" et leurs semblables, en fixant les cadres contextuels qui guident l'usage : contexte de justification et contexte d'explication.

Au total 110 étudiants universitaires de la région de Tokyo (46 hommes et 64 femmes), âgés de 19 ans à 29 ans ont participé à nos expériences. On leur a demandé de concevoir un dialogue succinct entre deux personnes très proches, de même sexe, pour deux contextes différents mentionnés ci-dessus. Des exemples réels, extraits d'une enquête, ont été montrés aux sujets comme dialogues-modèles [3]. Les consignes étaient les suivantes :

Contexte de justification :

L'objectif du script est de refuser une invitation en donnant une raison. Le dialogue doit comporter 4 étapes.
1 Vous êtes invité à boire un verre avec un/e ami/e.
2 vous déclinez l'invitation en précisant la raison.
3 votre ami/e la comprend.
4 vous vous séparez en vous saluant.

Contexte d'explication :

L'objectif est d'expliquer la situation par rapport à la remarque de votre ami sur votre santé. Le dialogue doit se dérouler comme suit :
1 On vous fait une remarque sur votre forme physique.
2 Vous en expliquez les raisons.
3 L'autre vous conseille de rentrer pour vous reposer.
4 Vous acquiescez et vous vous séparez.

2.2. *Mode parlé : Enregistrement des dialogues*

La seconde -expérience phonétique- consistait à enregistrer les dialogues contenant la même sorte de particules finales, réalisées par les sujets des deux sexes.

Nous avons réalisé deux sortes d'enregistrement : lecture de dialogues et dialogues libres. Pour la lecture, nous avons utilisé deux scripts-exemples pour chaque contexte: Justification et Explication. Un exemple a été lu deux fois en permutant le rôle de chacun dans le "couple". Ces exemples-modèles sont neutres par la forme et le contenu du point de vue des différences de sexe, et exempts de marques pouvant faciliter la réalisation phonétique, telles que point d'interrogation ou signe d'allongement vocalique. Après lecture, les sujets se sont parlé librement deux fois dans chacun des deux contextes, là également en permutant les rôles. Pour une paire de sujets, nous avons donc eu huit versions de lecture et autant de dialogues libres, soit environ une demi-heure d'enregistrement. L'enregistrement a été effectué en deux endroits différents, dont une chambre anéchoique munie d'un microphone et d'un magnétophone digital, ainsi que dans une pièce calme utilisant un micro à casque connecté sur un enregistreur digital. Parmi les sujets participant à l'expérience de composition des dialogues, 26 femmes et 20 hommes, au total 46 personnes, ont prêté leur voix. Nous avons transcrit leurs dialogues libres avant toute analyse.

3. ANALYSE DES RÉSULTATS

3. 1. Mode écrit

Pour notre première question de recherche, nous nous intéressons plus particulièrement à la forme linguistique placée en fin d'énoncé qui concerne l'expression de justification et d'explication dans les scripts composés et transcrits. Parmi les 110 sujets, quelques-uns répètent deux fois la même forme ou la réitèrent sous une forme différente, ce qui donne un total de 119 réponses pour le contexte de justification et 131 pour celui d'explication.

Le tableau 1 récapitule ces réponses selon le mode de production : écrit/parlé, le contexte : justification / explication, le sexe : femme/homme. Nous avons retenu six formes finales, les plus fréquentes : "~nda", "~no", "~kara","~te/de", "~shi" et "~kamo".

Nous avons examiné statistiquement ces données de fréquence : test de chi-2, test de Kruscal-Wallis et test de Friedman. Les facteurs expérimentaux ne montrent aucune différence significative dans les oppositions de mode: écrit/parlé, contexte: justification/explication, et le sexe: homme/femme. Pour les 3 formes finales utilisées dans nos exemples modèles: "~nda", "~kara" et "~te", la valeur calculée de chi-2 n'ayant pas dépassé le seuil critique à 5%, voire 10%, la différence entre homme et femme s'est révélée nulle.

	Ecrit : Justification		Ecrit: Explication	
	Homme	Femme	Homme	Femme
"-nda"	11	34	18	33
"-no"	0	9	0	1
"-kara"	7	3	6	1
"-te/de"	1	3	12	27
"-shi"	3	0	1	3
"-kamo"	0	0	2	6
	Parlé: Justification		Parlé: Explication	
	Homme	Femme	Homme	Femme
"-nda"	20	18	17	20
"-no"	0	4	1	0
"-kara"	7	4	4	2
"-te/de"	2	7	10	17
"-shi"	2	0	1	0
"-kamo"	0	2	0	3

Table 1 : Fréquence des formes terminales observées.

Dans nos données, la forme finale "~nda" historiquement de style masculin, s'emploie sans distinction de sexe autant dans la composition des dialogues que dans la réalisation phonétique des dialogues libres. En effet dans le contexte de justification, plus de la moitié des sujets féminins (52%) ont utilisé "~nda". D'autre part, la fréquence de "~no" dit de style féminin est singulièrement faible chez les sujets féminins. De ce fait, l'opposition "~nda" vs "~no" ne reflèterait plus la différence de style entre homme et femme. Si la conception de dialogue permet de sonder la conscience linguistique des sujets, la parole réalisée en dialogue permet de vérifier la réalité linguistique de la situation actuellement en cours. Le passage de "~no" vers "~nda" se dérouleraît, pensons-nous alors, dans la profondeur inconsciente.

3. 2. Mode parlé

Prétraitement des données

Nous nous intéressons a priori à la partie finale de l'énoncé, partie incontestablement plus riche en caractéristiques typiques prosodiques. Concrètement, nous avons mesuré la fréquence fondamentale (dorénavant F0) et la durée sur les deux dernières syllabes de l'énoncé, qui correspond à l'emplacement de la forme finale. Ainsi nous pouvons obtenir une indication potentielle de l'intonation et du rythme pour une phase ultérieure d'étude.

Pour déterminer la hauteur perceptible de la F0, nous avons tenu compte des résultats expérimentaux du glissando de la parole [6], et nous avons interpolé les points expérimentaux discrets par la fonction de Gompertz ci-dessous (1) afin de couvrir l'intervalle temporel allant

de 50 msec jusqu'à 400 msec.

$$\log_e G(x) = \log_e 13.35 + 0.67^{0.04x-1} \log_e 3.13 \quad (1)$$

Si le rapport entre les points de départ et d'arrivée d'une montée dépasse le seuil calculé **G** pour la durée du segment (**x** en ms.), la montée peut être perçue en tant que telle. De plus, même si cette montée est perçue, l'oreille ne l'intègre pas dans sa totalité ; le système auditif ne suit que jusqu'à la hauteur située aux 2/3 de la variation [7]. Dans le cas de la variation infraliminaire au seuil, un ton statique peut être perçu avec une hauteur correspondant également aux 2/3.

Dans le cas d'une variation bidirectionnelle, « montant – descendant » ou « descendant – montant », si elle est perceptible, la première partie de la variation est perçue jusqu'au point d'inflexion, et la deuxième partie est tronquée en hauteur aux 2/3. Ces règles d'interprétation de la F0 ont été appliquées au segment final de l'énoncé à examiner. Si tout le segment est entièrement voisé, nous l'avons traité comme une seule unité tonale. En revanche, si la courbe de F0 est interrompue par une consonne sourde au milieu, nos règles d'interprétation se sont appliquées à chaque partie voisée.

Statistiques des données acoustiques

Nous avons effectué les analyses de variance (dorénavant ANOVA) uniquement sur les données de lecture, avec la F0 et la durée comme variable dépendante. Les facteurs expérimentaux à prendre en compte, avec chacun deux niveaux, sont le sexe et le contexte. Le facteur des exemples a été considéré comme celui de répétition. Les données de F0 pour l'ANOVA, la valeur mesurée en Hz a été convertie en ***cent signé*** afin de neutraliser la différence inhérente au sexe du registre de la voix.

A. Résultats de l'ANOVA : F0

Le facteur du sexe s'est révélé hautement significatif ($F_{(1,148)} = 19.825$, $P < 0.0001$), de même celui du contexte ($F_{(1,148)} = 7.864$, $P<0.01$).

B. Résultats de l'ANOVA : durée.

En prenant le rapport de durée entre les deux dernières syllabes comme variable dépendante et les mêmes facteurs expérimentaux, nous avons procédé à une autre ANOVA. Le facteur sexe a montré un effet remarquablement significatif ($F_{(1,148)} = 20.003$, $P<0.00001$), mais l'effet pour le facteur contexte n'a pas été significatif. Toutefois, l'interaction de ces deux facteurs a révélé un effet statistique significatif ($F_{(3,148)} = 6.615$, $P<0.05$).

Interprétation phonétique

Si nous examinons ensemble les résultats statistiques sur la F0 et la durée, nous remarquons que les sujets féminins montrent une plus ample variation dans ces deux paramètres acoustiques. Les hommes répètent un schéma intonatif descendant et ne varient pas énormément en tempo. Regardons ces phénomènes de près en nous référant à nos schémas.

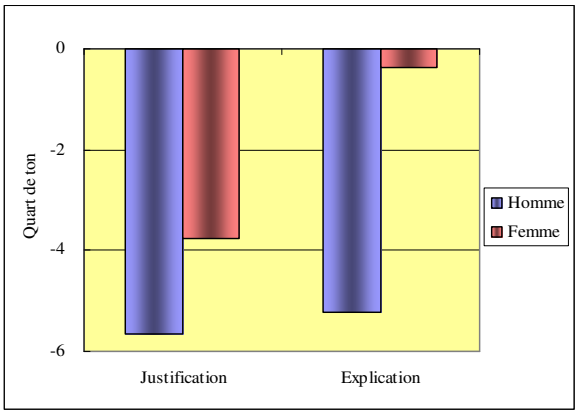


Figure 1 : Chute mélodique en position finale

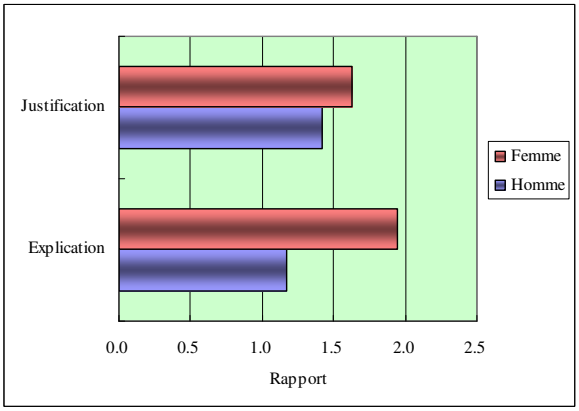


Figure 2 : Allongement de durée en position finale.

Sur le schéma de la Figure 1, les deux contextes confondus, on constate que la voix masculine descend en note musicale (1.36 de moyenne) et s'allonge de 29% environ sur la syllabe finale suivant la figure 2. En revanche, les sujets féminins utilisent une chute légère qui n'atteint pas un ton, et elles allongent d'en moyenne 80% de plus la dernière syllabe. Il est à noter que dans le contexte d'explication, la durée finale double et ce, avec une hauteur non descendante. Ce qui fait ressortir que dans ce contexte, elles utilisent une intonation à mi-hauteur traînante.

Le contraste d'allongement final attire une attention particulière par rapport au seuil différentiel de durée. D'après les travaux effectués dans ce domaine [8], un allongement de 29% dépasse à peine le seuil différentiel entre deux sons. En revanche, le dédoublement de la durée ne se produit que dans une opposition phonémique du type voyelle simple vs. voyelle longue, consonne simple vs. consonne géminée, par exemple [9].

4. ANALYSE DE GRAPHIE

Dans les dialogues composés, nous avons observé une sorte de personnalisation de l'expressivité en ayant recours aux symboles orthographiques. L'effet recherché est réalisé par les points de suspension pour l'hésitation, par des voyelles dédoublées ou suivies d'un trait indiquant un allongement vocalique. Cette représentation orthographique non authentique semble un reflet de leur conscience linguistique pour une phonétisation interne ou sous-jacente du texte. Il peut y avoir une différence selon le sexe de la personne qui écrit.

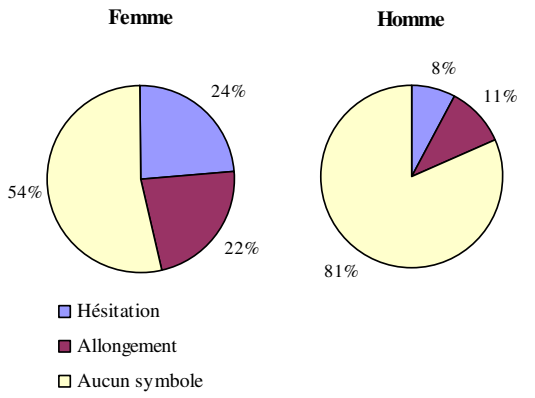


Figure 3 : Symboles spéciaux utilisés.

Dans la figure, nous représentons les trois catégories de spécifications : hésitation, allongement, et non recours à ceux-ci. Indifféremment aux contextes, ce sont les sujets féminins qui renforcent leur intention communicative par ces moyens, et presque la moitié des femmes y recourent. En revanche les hommes ne l'utilisent que 19 fois dans 103 dialogues.

Les tests de Chi-2 ont montré une différence significative dans chacun des contextes. Justification :$\chi^2 = 8.4246$, $\chi_{.01}2$ (df = 2) = 6.6349 ; Explication : $\chi^2 = 12.9954$, $\chi_{.001}2$ (df = 2) = 10.8), et bien entendu, avec les deux contextes confondus, la différence entre hommes et femmes s'est avérée significative ($\chi_{.001}2$ (df = 2) = 10.8). En revanche, aucun effet statistique n'est constaté entre les deux classes de symboles.

Nous n'avons donné aucune indication d'oralisation éventuelle pour la tâche de composition, les sujets féminins étaient conscients du rendu phonétique. Ce qui va dans le même sens que dans l'examen des caractéristiques prosodiques décrites ci-dessus.

5. CONCLUSIONS

Nous avons effectué une série d'expériences sur les différences linguistiques selon le sexe des personnes impliquées dans un dialogue informel entre amis, dans des contextes de justification et d'explication. Nous n'avons trouvé aucune différence significative dans le choix des formes linguistiques de fin d'énoncé, mais les sujets féminins montrent une variation plus contrastée dans l'intonation et le rythme. De plus, même dans l'écrit de dialogues, elles recherchent une oralisation inconsciente.

BIBLIOGRAPHIE

[1] M. Kobayashi. Sedai to joseigo. *Nihongogaku*, Meiji-shoin, Tokyo, volume 12-5, pages 181-192, 1993.

[2] Y. Ozaki. Josei senyô no bummatsu-keishiki no ima. In *Josei no kotoba*, Hitsuji-shobô, Tokyo, pages 33-58, 1997.

[3] H. Yabe. Hanashi-kotoba ni okeru danjosa to shite mita 'nda'. In *Nihon to chûgoku kotoba no kakehashi*, Kuroshio-shuppan, Tokyo, pages 187-196, 2000.

[4] N. H. McGloin. Shûjoshi. *Nihongogaku*, Meiji-shoin, Tokyo, volume 12-6, pages 120-124, 1993.

[5] C. Suzuki. Bummatsu-hyôgen no intonêshon to danjosa. *Kotoba*, Gendai-Nihongo-Kenkyûkai, Tokyo, volume 20, 1999.

[6] M. Rossi. Seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica* 23 : 1-33, 1971.

[7] M. Rossi and M. Chafcouloff. Recherche sur le seuil différentiel de fréquence fondamentale dans la parole. *Travaux de l'Institut de Phonétique d'Aix*, Aix-en-Provence, volume 1, pages 179-185, 1972.

[8] D. H. Klatt and W. E. Cooper. Perception of segment duration in sentence contexts. In *Structure and Process in Speech Perception*, A. Cohen ; S. G. Nooteboom (Eds), Springer, New York, pages 69-89, 1975.

[9] Y. Nishinuma, D. Duez and C. Paboudjian. Duration of consonant clusters in French : Automatic classification rules. In *Proc. Euro speech '89*, pages 260-263, 1989.

REMERCIEMENTS

Une partie de nos travaux a été financée par : *Awards for Significant Research Projects (Jûten-kenkyûhi, H-17),* Université Tokyo Gakugei, ainsi que *Grant for Specific Research Issues (Tokutei-kadai-kenkyûhi, 06/4-08/3),* Faculte des lettres, Université Chuo, Tokyo.

Session V

Conférence Invitée

Mardi 13 juin 2006 - 09h00 10h00

Analyse de la violence verbale : quelques principes méthodologiques

Claudine Moïse Université d'Avignon

Claudine.moise@univ-avignon.fr

Remarque. Ce texte est une synthèse du travail que nous menons et des articles que nous avons écrits avec mon équipe de recherche, notamment Nathalie Auger, Béatrice Fracchiolla Christina Schultz-Romain. Le projet de recherche, *La violence verbale*, a été financé par La Délégation Interministérielle à la Ville et la Région Provence Alpes Côte d'Azur.

ABSTRACT

The official discourses in France for describing the social and linguistic tensions use frequently the term of « violence verbale ». Thus the « violence verbale » represents insults, threats identified by particular lexical uses of language. But, it is never analysed in terms of social norms. Based on different interactions data collected in public and private areas and on film scripts, this presentation will describe the emerging of the « violence verbale ». We will first focus on the different uses and values attributed to linguistic and social norms in the interactions. But then, we will explain how the use of « violence verbale » in public situations allows the reproduction of social order and how the transgressions of norms with the insult define linguistic uses to transform the « rapport de forces symboliques ».

1. INTRODUCTION

Si l'on peut décrire la violence verbale à partir d'actes de langage repérables et analysables (*insulte, mépris, dénigrement, menace*, etc…), dans une perspective pragmatique (Lagorgette, D. Ogier, C. Rosier, L.), voire d'analyse conversationnelle (Traverso, V., Kerbrat-Orecchioni, C., Vincent, D.), la violence verbale doit être aussi appréhendée dans sa globalité (et au-delà du fait linguistique). Les pratiques discursives et les interactions renseignent sur les pratiques sociales, c'est-à-dire que, par les jeux utilisés dans l'interaction, les locuteurs donnent sens à leurs actions, aux prises de pouvoir et aux positionnements sociaux (Gumperz, J. Martin-Jones, M., Gal S., Heller, M.). Les formes adoptées dans l'interaction violente peuvent donc être comprises comme des actes individuels, adhésion ou distanciation par rapport à l'interlocuteur et mais aussi comme des actes socialement inscrits, signes d'identification, d'appartenance ou de résistance.

Nous avons défini la violence verbale comme des « montées en tension interactionnelle» marquées par des « déclencheurs » spécifiques, processus qui s'inscrit dans des rapports de domination entre les locuteurs, des télescopages de normes et de rituels, des constructions identitaires. Il s'agit alors de dire une sociolinguistique du sujet, en œuvre dans les interactions violentes.

Je montrerai dans un premier temps comment, d'un point de vue méthodologique, la construction de l'analyse est sujette aux conditions d'élaboration et à la diversité des situations et des corpus. Ensuite, à partir d'un corpus cinématographique, je rendrai compte du processus de montée en tension dans une perspective intersubjective. De cette façon, la violence verbale se structure en étapes séquentielles où contextes d'énonciation et signifiances culturelles et relationnelles jouent un rôle essentiel. Enfin, je m'attacherai à décrire un phénomène d'acte de langage signifiant de la violence verbale, l'insulte.

2. PRINCIPES METHOLODOLOGIQUES

2.1. Ancrage sociolinguistique

J'ai, à diverses reprises (Moïse [14][15]), posé mes « intentions sociolinguistiques ». Je ne reviendrai pas ici sur les différents champs de la sociolinguistique relatifs à l'histoire même de la discipline si ce n'est pour souligner à la fois combien ils sont divers, allant du variationnisme qui s'est emparé de l'étiquette « sociolinguistique » (Chambers [5]), à celui plus mouvant d'une sociologie du langage. Dans un sens, il est une première sociolinguistique qui s'intéresse à la société pour ce qu'elle nous dit sur la langue. C'est prendre souvent les différences sociales à travers des catégories préétablies, essentialistes (sexe, âge, origine, catégorie socio-professionnelle) dans une forme de réduction nécessaire, maniable et pratique, et s'en servir pour lire les variations en langue. La sociolinguistique française, marquée par l'histoire de la discipline et donc par la linguistique structuraliste, s'est longtemps, par frilosité aussi, limitée à une dimension descriptive. L'autre sociolinguistique (ou les autres sociolinguistiques) qui m'intéressent davantage disent la société à travers l'étude de la langue, des pratiques et des discours. Serait-ce une sociologie du langage, terminologie très (trop) associée à J. Fishman dès les années 60, critiquée aujourd'hui

d'ailleurs (Williams [20]) ou une anthropologie linguistique ? Mais cette autre sociolinguistique englobe alors un champ vaste, plus significatif dans le monde anglo-saxon (Mesthrie [13]), allant de l'analyse des discours en œuvre dans la société à l'analyse des interactions notamment, voire à une nouvelle façon d'aborder la variation comme ressource en contexte (Gadet [6]). C'est, de façon globale, cette sociolinguistique qui constitue mon cadre de recherche.

2.2. Méthodologie de l'enquête

Les différents terrains

Je considère le terrain, au-delà du site, et avec mes collègues sociolinguistes, de Didier de Robillard à Cyril Trimaille, comme des matériaux empiriques et des réseaux de relation à jamais circonscrits, ou alors délimités pour les besoins humains de l'enquête. Dans une telle optique, on ne pourra être qu'en exploration, à jamais insatisfaits, en prise avec une réalité que nous participons à construire.

Nous avons mené une première enquête dans un collège de Zone d'Education Prioritaire à Perpignan dans le sud de la France. L'étude portait sur la violence verbale dans le cadre institutionnel et scolaire. Il s'est agi d'analyser ses effets dans des rapports asymétriques. En ce sens, si nous avons fait quelques observations concernant les comportements entre élèves, notre attention s'est portée avant tout sur les relations enseignants, principal et personnel administratif face aux élèves.

Par la suite, nous avons envisagé de travailler à partir de corpus plus diversifiés, notamment des enregistrements en milieu informel (conversations familières, interactions dans les transports en commun, etc) et des extraits de scénarios de films.

Actuellement, je m'attache davantage à cerner la notion de *privé/public* et de voir en quoi elle joue dans la structuration de la violence verbale. Que dire de l'espace de la rue, de celui de la voiture, du guichet de poste ? Quels rôles jouent les places, dissymétriques ou symétriques, des locuteurs en situation de violence verbale ? Quels sont alors les schémas de la violence verbale ? Chaque situation d'interactions renvoie à des catégories sociales spécifiques et opérationnelles qui modulent les formes de violence verbale.

La subjectivité du chercheur

Finalement, selon une démarche caractéristique en ethnographie de la communication, l'enquêteur est en position d'observateur, distant toutefois du groupe à observer n'en détenant pas lui-même les codes. Il utilise prises de notes, consultation de tout document produit par les institutions notamment, enregistrements de moments de tensions, de conversations authentiques, spontanées mais aussi d'interviews semi-dirigées, rendant compte des jugements, représentations, explications des locuteurs sur leur propre production langagière. En même temps, dans une perspective ethnométhodologique, la construction sociale n'est pas une donnée stable et préexistante, indépendante des interactions sociales, elle se modèle et module au gré des actions humaines, des savoirs, des apprentissages et des croyances de chacun. En examinant les comportements de l'intérieur, en faisant une observation attentive des processus par lesquels les acteurs interprètent constamment la réalité et réinventent la vie, on peut dresser des configurations sociales et une production singulière des interactions.

En ce sens, le chercheur lui-même, dans une démarche empirico-inductive, est partie prenante des situations d'interactions et sa propre subjectivité participe à la fois de la configuration du réel et de l'analyse. Face à une telle recherche, le chercheur se trouve saisi par diverses contraintes.

Dans les situations de tensions et de violence verbale, le chercheur pourra prendre diverses positions selon le contexte d'énonciation, rôle de témoin, de médiateur, d'observateur, de provocateur, d'acteur de la violence verbale. En milieu formel et institutionnel (milieu scolaire ou public), il est avant tout témoin voire médiateur.

En milieu informel, il est souvent difficile de saisir promptement son matériel d'enregistrement. Le mieux, surtout si l'on est soi-même partie prenante de la violence verbale, qu'on veuille ou non entrer dans le jeu interactionnel, est de faire confiance à sa mémoire pour tenter de retranscrire au mieux par la suite les interactions. Parfois, il est possible d'enregistrer mais il manque souvent la première phase de montée en tension. Il s'agit aussi de contrôler ses émotions et sa propre violence intérieure. Un tel projet, qui permet aussi de mener des interventions de formation dans un cadre de recherche action, demande une capacité réflexive du chercheur face à ses comportements et pratiques. Dans une telle perspective, notre équipe de recherche a dû se former à des pratiques de « gestion de conflits » ou de « communication non violente », reposant, au-delà de l'analyse et de la compréhension des phénomènes langagiers, sur une appréhension psychologique des relations verbales.

Si le chercheur n'est pas partie prenante, par sa présence même et notamment en mieux formel et scolaire, il lui est demandé souvent d'être témoin voire médiateur. Un des procédés fondamentaux mis en œuvre par l'école de Palo Alto (Watzlawick, Weakland, et Fisch [21]), en particulier, celui qui consiste à régler un problème en sortant du cadre de la pensée auquel nous sommes généralement habitués : pour rompre certains schémas de communication, il faut ainsi s'extraire de la binarité du rapport, prendre de la distance et aller, d'un point de vue géométrique, chercher la tierce solution à l'extérieur. Plus simplement, il s'agit d'insérer un point de vue tiers

dans une relation binaire. L'idée d'une psychosociolinguistique (Van Hooland [19]) requiert la présence d'un tiers dans la communication - la partie " psycho "- qui joue un rôle mixte de réflecteur, mitigeur, mise à distance et objectivation par comparaison ; qui interroge du moins ce qui est *donné pour réel* par le locuteur à l'interlocuteur. Si l'on comprend la notion de sociolinguistique relativement à la dimension sociale, spatiale et temporelle, on peut comprendre alors la dimension " psycho " dans le sens d'une nécessaire présence médiatrice (physiquement présente ou non) entre les deux individus qui élaborent un type d'échange fonctionnant jusque-là en cercle binaire et fermé. Cette idée est intéressante dans la mesure où elle réintroduit l'idée d'un triangle correspondant à la réalité pronominale et grammaticale (je, tu, il/elle) en redonnant une véritable place de personne au " il/elle ", alors que pour les interactions verbales, on s'attache(ait ?) essentiellement à la relation binaire, entre " je " et " tu " en traitant le " il "/ "elle " comme une " non-personne ". Il est probable qu'une telle perspective annonce un changement à l'œuvre en ce qui concerne les représentations de l'altérité *à travers de nouvelles formes d'interactions.* On peut remarquer à cet égard l'importance donné ces dernières années au rôle (nouveau) du *médiateur*, du *négociateur*, qui apparaît dans la résolution non violente des conflits (Auger, Fracchiolla, Moïse [1])

2.3. Méthodologie de l'analyse

La « violence verbale » est plus à définir comme des effets de rupture dans les interactions et de montée en tension, rendant impossible toute action de négociation. La violence verbale est à saisir comme un processus langagier dynamique qui se rejoue sans cesse dans les prises de parole. Elle se construit alors dans un double mouvement. Elle est signifiée d'un côté par le contexte d'énonciation. Le contexte est à saisir, non seulement au sens énonciatif (le co-texte) mais au sens plus large - social, ethnique, symbolique - comme une composante nécessaire et incontournable de la construction du sens (Boutet [3]). Les rites d'interaction en jeu répondent à ce contexte, et les injures, insultes (ou perçues comme telles), ne sont finalement que la partie visible, repérable de l'agression. D'un autre côté, elle est signifiante des rapports sociaux, acte social en elle-même et construit, participe, bouleverse les configurations établies d'un certain ordre légitimé.

Ainsi, notre approche repose sur plusieurs niveaux d'analyse, que ce soit celui des actes de langage dans une perspective performative, celui de l'énonciation ou de l'ethnographie de la communication et de l'analyse des conversations. On utilisera donc à la fois les principes de l'analyse conversationnelle (Sacks, Traverso), de l'analyse discursive (construire des « espaces discursifs » –Heller - et observer leur circulation – Vincent, Rosier-), des théories des actes langage (Austin, Kerbrat), de l'analyse sociolinguistique du style (Gumperz, Gadet, Kallmeyer). Les marquages stylistiques fonctionnent comme moyens de contextualisation donnant des cadres de référence contribuant au sens des activités communicatives. Les formes de style fonctionnent comme des symboles d'identité qui créent de l'adhésion sociale à travers des cadres sociaux, marquent des différences, permettent de se positionner dans la société, sur les marchés sociaux et sur les arènes publiques.

3. LA MONTEE EN TENSION

3.1 Mécanisme de la montée en tension

Processus

La violence verbale est inhérente au conflit qui est une divergence de points de vue, manifestes sur le plan interpersonnel et des normes sociales (il peut y avoir par exemple divergence sur la notion de « respect » liée aux nuisances sonores de voisinage) et qui entraîne une forte tension entre les locuteurs. La violence verbale est d'autant plus « radicale » qu'elle s'inscrit dans une opposition caractérisée entre les interlocuteurs. Nous avons donc dégagé, suite à différentes analyses de situations de violence verbale, les étapes suivantes constituant une *montée en tension* dans les conversations.

1ère étape : la violence potentielle

La violence verbale est à relier au contexte général de communication, forme de *climat général* (Galatalo, Mizzau [7]). On parlera de *violence potentielle*, liée à la personne elle-même, à son agressivité comportementale, ou liée à un contexte supposé violent, construit à travers représentations ou mises en scènes médiatiques, comme sont les images renvoyées des banlieues.

2ème étape : la violence embryonnaire ou amorce de la violence verbale

Comme nous l'avons montré par ailleurs, il est des éléments identifiables linguistiquement d'une amorce de la violence verbale. On peut noter parmi eux, *l'agressivité* avec changement prosodique et posture particulière du corps, *le harcèlement verbal* avec répétition interactionnelle dans différentes séquences conversationnelles, *les joutes verbales* caractérisées par des changements de registres verbaux. A ranger dans ces figures, toutes sujettes à l'intersubjectivité des locuteurs, *l'impolitesse* et *l'incivilité.* On considère *l'impolitesse* comme une rupture des rituels conversationnels interpersonnels (refus de dire bonjour) tandis que *l'incivilité* serait à prendre d'un point de vue des codes sociaux (utilisation du téléphone portable dans le train par exemple).

L'amorce de la violence verbale est « lancée » par un locuteur A et va entraîner certains modes de réactions de la part du locuteur B.

3ème étape : la violence cristallisée

Face aux attaques de A, le locuteur B peut adopter différents comportements, notamment entrer résolument dans le conflit et prendre part à la montée en tension. Dans ce cas-là, il est fait usage de *l'insulte*, de la *menace* (souvent dans une forme d'injonction physique, « je vais te casser la tête »), du *mépris*, actes pragmatiques repérables dans le discours à forte valeur perlocutoire (Moïse [16]). La montée en tension se joue et se rejoue dans les différentes prises de parole des locuteurs sous formes de boucles interactionnelles ou A et B interchangent leur place dans une joute verbale.

Cette entrée marquée dans la violence verbale peut être dépassée ou évitée à travers d'autres résolutions conversationnelles. Le locuteur B peut tenter de mettre un terme au conflit par la négociation, qui portera sur l'objet même du conflit ou sur la relation interpersonnelle (Ott, [17]). Dans ce cas, il faut que les deux locuteurs soient capables de « s'entendre » hors de tout sentiment d'atteinte à la face. D'une autre façon le locuteur B peut décider d'opter pour la fuite ou l'évitement. La fuite consiste à se taire, voire à physiquement partir, ou à opter pour un changement thématique (« bon parlons d'autre chose »). L'évitement consiste à rester dans la thématique sans contre-attaquer, comme peuvent l'être des marques d'humour. Véronique Traverso [18] parle pour l'évitement de *dispute évitée* quand il y a désaccord sans négociation ni explicitation, chacun des locuteurs restant sur ses positions ; on est dans un échange immobile.

4ème étape : la violence physique

Au-delà de la violence verbale, l'ultime recours pour se faire entendre est la violence physique dans une forme de passage à l'acte souvent annoncé pragmatiquement – par la parole ou le mimo-gestuel - dans les montées en tension (« si tu continues, ça va mal se passer »).

Analyse du corpus

Pour plus de clarté dans la démonstration, j'ai choisi ici pour l'analyse un extrait du film *Karnaval* de Thomas Vincent. Dans la séquence d'ouverture du film, on assiste à un conflit entre un père et son fils. Le père travaille avec ses deux fils dans un garage à Dunkerque.

Première scène dialoguée.*Le père sort du garage avec Nasser, le fils aîné ; ils sont de dos, le fils cadet les suit de très près, derrière eux, de dos aussi.*

*1.**Fils cadet***. David [*geste du bras à l'adresse du père alors que celui-ci est toujours de dos*] lui à Marseille lui

*2.**Père***. quoi / un garage normal c'est pas un garage normal ici [*il lève les bras au ciel*]

3. ***Fils cadet***. David [*geste du bras à l'adresse du Père*] à Marseille il (se) fait cinq mille sept par mois cinq mille sept

*4.**Père.*** et vas-y à Marseille tu verras comment c'est Marseille

5. ***Fils cadet.*** de toute façon toi tu préfères faire toujours tout / plutôt que de me payer / Nasser lui là tu le payes bien lui [*désignant Nasser, toujours de dos*]

6. ***Père***. Nasser Nasser [*ratage*] [*il le montre*] / il travaille lui

7. ***Fils cadet***. Nasser Nasser [*il le montre*] / il te lèche le cul / Nasser c'est tout ///

[*Le Père et Nasser se retournent*]

8. **Père.** quoi // tu me dis ça à moi ton père // [*Le père gifle le fils et maugrée*]

[*La caméra est maintenant derrière le Père et Nasser ; le fils se retrouve donc de face*]

9. ***Fils cadet au père***. c'est la dernière fois que tu me touches

*10.**Fils aîné, Nasser***. allez allez [*il s'interpose*]

11. ***Fils cadet au père.*** la prochaine fois je t'envoie à l'hôpital moi

12. ***Fils aîné. Nasser***. arrête tes conneries / allez arrête

13. ***Fils cadet.*** [*ratage*] / vas-y toi / protège-le toi [*il montre le père*]

14. ***Fils aîné. Nasser***. tu lèves la main sur papa maintenant /

15. ***Fils cadet***. un père ça / un père ça / un fils [*il s'auto-désigne*]/ moi je suis un esclave ici moi

Conventions de transcription

• Les pauses, selon leur durée, sont marquées par /, ou //, ou encore ///.

• Un mot incompréhensible se note par (X), un passage plus long par (XXX), une incertitude de transcription par (de X).

• L'allongement est noté par :

•*(rires)* est un commentaire d'un comportement non verbal.

• L'hésitation entre deux formes, bien souvent morphologiques, est citée entre parenthèses (j'ai été / j'étais) (i regarde(nt) (ces / ses). Cette hésitation peut se manifester aussi entre la forme pleine et sa non manifestation (ça a été / ça ø été)

• Les paroles simultanées sont soulignées

• Les liaisons non conformes à la norme sont marquées avec trait d'union, *j'suis-t-allé.* Le *n'* de liaison ou de négation est marqué entre parenthèse, *on (n') y était pas* .

1[ère] étape, la violence potentielle. S'il est difficile de juger d'une violence potentielle chez les locuteurs, puisqu'il s'agit de la première séquence du film et que le spectateur ne sait encore rien de la psychologie des personnages, en revanche, le décor instaure un climat général de malaise. La scène se déroule dans une zone industrielle qui semble laissée à l'abandon. Les protagonistes déambulent en bleu de travail dans ce paysage urbain interlope qui renvoie à des représentations médiatiques de la violence.

2[ème] étape, la violence embryonnaire. Au tout début, le fils (locuteur A) agresse le père et on retrouve des procédés langagiers, caractéristiques de la violence verbale. Le fils cadet suit son père et son frère de très près et manifeste ainsi un certain harcèlement physique. En même temps, il fait preuve de harcèlement par la parole : tours 1 et 3, « David à Marseille », tour 3, « cinq mille sept ». Ses propos sont agressifs car signifiés par une prosodie montante et une accélération du débit de parole en fin de segment, matérialisées dans la transcription par les nombreuses flèches montantes.

3[ème] étape, la violence cristallisée. Le père (locuteur B) tente la fuite, en tout cas physiquement. Il refuse de se tourner vers son fils pour parler. Dans un premier temps, il essaie alors, contraint par son fils, (tour 2) de s'expliquer pour se justifier. Par sa question rhétorique (tour 2), il cherche à montrer que sa façon de travailler peut être légitime (« normal »). Malgré la tension, il offre une voie de négociation autour de cette question de ce qu'est la « normalité » dans le fonctionnement d'un garage. Cette négociation ne pourra se réaliser car au tour 3, en réponse, le fils signifie son désaccord en avançant l'argument du salaire. Pour lui, dans un garage « normal », on devrait gagner l'équivalent du salaire minimun (SMIC à *cinq mille sept* cent francs). Le père reste lui aussi sur ses positions (tour 4) en lui montrant qu'à Marseille, le traitement sera sans doute le même. Entre les tours 2 et 4, il y a eu échec d'une éventuelle négociation et nous sommes dans une cristallisation de la violence verbale. Déjà dans le tour 4, les injonctions du père (« vas-y à Marseille ») expriment le rejet à la fois de l'argument et de la personne, forme de déconsidération et de mépris de l'autre, sans respect de sa face. La réponse du fils (tour 5) remet en cause le comportement de son père par une expression généralisante sur l'autre, marquée par les adverbes « toujours », « tout »(« de toute façon toi tu préfères toujours tout »), tout en lançant un argument à négocier, la différence de traitement entre les deux frères. Au tour 6, le père ne discute pas sur le fond, c'est-à-dire sur la considération des deux frères, mais reprendra davantage sur le registre méprisant (« il travaille lui »). Ces séquences interactionnelles rebondissent les unes par rapport aux autres, font effet d'accumulation, et expriment la montée en tension. Des tours 4 à 6, les locuteurs sont dans le mépris de l'autre qui va se conclure dans l'insulte, en forme de dernier recours (tour 7) pour évacuer le trop de tension (« il te lèche le cul / Nasser c'est tout »).

4[ème] étape, la violence physique. L'insulte est souvent le point de rupture avant la violence physique. Le père (tour 8) ne peut accepter l'irrespect manifesté par son fils et pour réparer cet acte menaçant répond par la gifle, dernière issue possible pour retrouver la face (« quoi // tu me dis ça à moi ton père »). Sans l'interposition du frère aîné, l'altercation aurait pu s'envenimer (tour 11) et le fils cadet est réduit à la profération d'une menace physique (« la prochaine fois je t'envoie à l'hôpital »).

3.2. Les déclencheurs de conflits

Pour comprendre comment se met en place la violence verbale, c'est-à-dire une certaine montée en tension conversationnelle entre les locuteurs, il faut saisir ce que nous avons appelé les *déclencheurs de conflits.* Nous avons pu observer qu'ils étaient liés à différents aspects des conflits, liés aux relations interpersonnelles, structurelles ou culturelles.

Le conflit de valeurs ou *culturel* (Ott, [17]). Les locuteurs sont en mésentente voire en opposition idéologique sur des représentations, des idées morales liées aux groupes sociaux ou ethniques. En ce sens, il va légitimer les différences et pourra s'orienter vers un conflit interpersonnel. Les différentes appréhensions des *universaux culturels* – relation à la famille, à la mort, au travail…- peuvent entraîner, dans des situations extrêmes d'incompréhension, des situations de conflit.

Le conflit structurel s'actualise dans la transgression des normes sociales qui maintiennent l'ordre établi et qui sont donc particulièrement identifiables dans les structures institutionnelles, système scolaire, entreprises, etc. Dans ces situations, les contrats de parole (rituels conversationnels implicites liés à un contexte particulier) sont malmenés par la violence verbale, rupture provoquée et engendrée par des sentiments d'injustice et de relations dominants/dominés (Auger, Fillol, Lopez, Moïse [2]).

Le conflit interpersonnel repose sur une remise en question de l'autre, dans un reproche de ce qu'il est, forme de *conflit d'identité* (Ott [17]) On est dans le non-respect de la face, sans précaution locutoire et dans une recherche de l'avantage conversationnel pour un maintien ou prise d'une place haute.

Analyse du corpus

Dans ce corpus, les trois aspects, déclencheurs de conflit sont visibles et interdépendants, même s'ils apparaissent à différents moments du désaccord. Le fils cadet remet en cause le fonctionnement d'une

« institution », le garage familial et la hiérarchie entre les employés – les fils - instituée par le père car à travail supposé égal, le traitement est différent (tours 1-3- 5). Cette situation entraîne un sentiment d'injustice où sont contestés les rapports de dominants/dominés et la notion de « normalité », c'est-à-dire la relation à la norme et à l'ordre préétabli. Dans un tel cas, le contrat de parole s'inscrit dans un contexte professionnel. Il aurait fallu pour le respecter demander un rendez-vous au « chef d'entreprise ». Mais, ici, à la dimension structurelle se rajoute la dimension personnelle puisque le patron est aussi le père. À partir du tour 5, le conflit s'actualise sous une forme interpersonnelle. On quitte l'objet du conflit pour passer à un conflit d'identité – remise en question de la personne et reproches. La remise en cause de l'attitude personnelle de l'autre passe par une accusation en « tu ». Il s'agit de la part du fils et du père de garder l'avantage soit par l'expression de la victimisation et de la culpabilisation (« Nasser lui tu le payes bien »), soit par celle d'un jugement (« Nasser / il travaille lui »). Or, les jugements interpersonnels sont ici liés à des représentations culturelles sur la famille et le travail. On peut admettre que le fils cadet remet en question l'entraide familiale au sein de l'entreprise, qu'il demande avant tout (tour 5) d'être considéré comme un employé à part entière (« tu préfères toujours tout plutôt que de me payer »), d'autant plus qu'il mérite considération en tant que fils (tour 15 «un père ça / un père ça / un fils / moi je suis un esclave ici moi »). On peut aussi faire l'hypothèse que la divergence de point de vue sur ces thèmes repose sur le conflit de génération et sur l'entraide communautaire et familiale due à l'histoire de l'immigration.

4. UN ACTE DE LANGAGE, L'INSULTE

Toute montée en tension se caractérise aussi par des actes de langage analysables d'un point de vue linguistique. En ce sens, nous nous sommes intéressés à des figures comme le malentendu, la menace, le harcèlement. Je voudrais ici me centrer sur l'insulte pour montrer combien elle participe de façon particulière de la violence verbale. Elle apparaît le plus souvent en situation dissymétrique comme dernier recours avec la confrontation physique. En situation informelle et symétrique, elle est plus fulgurance et joue de façon plus directe de la négation de l'autre.

4.1 Questions de terminologie

Du gros mot à l'injure

Souvent dans les dictionnaires, comme dans le *Petit Robert*, les choses sont mêlées ; « injure » est synonyme d'« invective », « insulte », « gros mot », « quolibet ». Effectivement. Mais c'est oublier que tout lexème peut prendre l'une ou l'autre valeur en contexte et, au-delà d'un sens, jouer sur des caractéristiques pragmatiques, essentielles pour ce qui est de la violence verbale. « Merde », quelle que soit la forme de « transgression langagière », peut être à la fois mot grossier, juron et injure, (Huston [8]) Le gros mot jouera sur la fonction référentielle du langage, fera référence donc à l'objet désigné (« la merde »). Le juron joue de la fonction expressive et sert le locuteur, pour ponctuer son discours, façon d'être dans l'emphase (« merde ! »). La transgression sera d'ordre scatologique, sexuel (« Putain ! ») ou sacré (« Nom de Dieu ! »). L'injure ou insulte vise l'interlocuteur dans une fonction impressive, « je te dis merde » ou même « tu es une merde », « espèce de merde ».

De l'injure à l'insulte

L'injure et l'insulte sont souvent employées l'une pour l'autre. Certains auteurs considèrent que l'insulte serait un jugement donné comme vrai, comme vérifiable sur l'interlocuteur et comme justifiable par le contexte. L'injure relèverait de l'imaginaire, du fantasme et de la provocation au-delà d'une vérité et d'un jugement vérifiable (Larguèche [11]). Traiter quelqu'un de « gros lard » s'il est gros relève de l'insulte, sinon de l'injure. Mais il est bien évident que les insultes ou injures reposent sur des jugements de valeur et donc sur des appréciations subjectives ; à partir de quelles normes pourra-t-on juger de la véracité des propos portés ?

L'injure comme l'insulte sont des actes de langage interlocutifs ; elles portent une force émotionnelle, voire pulsionnelle, et visent l'autre dans la volonté de le rabaisser et de le nier. Elles jouent un rôle éminemment perlocutoire (« Parce que je te traite de gros lard, tu vas te sentir comme ça »). Ce fonctionnement-là est rendu possible par des effets linguistiques. *Le trait axiologique est une propriété sémantique de certaines unités lexicales qui leur permet dans certaines circonstances de fonctionner pragmatiquement comme des injures, le marqueur illocutoire (effet sur l'interlocuteur) de l'injure étant la résultante complexe d'un ensemble de faits particuliers* (Kerbrat-Orecchioni [9]: 79). Parmi ces faits particuliers, il y a la valeur lexicale des lexèmes. Il semble évident que certains axiologiques négatifs puissent être réactivés pour faire injure : « gros » sera plus activé que « mince » par exemple. Il y a la forme syntaxique aussi, le terme péjoratif étant alors employé en fonction vocative avec notamment des modalités de catégorisation comme « espèce de ». Il y aussi l'intonation, tout terme, même neutre, pouvant se charger d'injure par la simple force d'évocation, façon de passer du constat à l'énoncé injurieux. En ce sens, l'injure aura toujours une force illocutoire, puisqu' un effet doit être produit sur l'interlocuteur. Mais pour que l'injure fonctionne pleinement, encore faut-il que l'interlocuteur la perçoive comme telle, en bref qu'elle touche, qu'elle déstabilise et non qu'elle conforte l'autre dans ses croyances. Comme le montre C. Kerbrat-Orecchioni [9], si je traite l'autre d'anarchiste « Anarchiste ! », et s'il répond « parfaitement », l'effet souhaité est rompu.

Finalement, injure ou insulte sont identifiables linguistiquement dans leur forme et visent l'autre dans un effet illocutoire voire perlocutoire dégradant et à travers une subjectivité partagée. Ainsi, les vannes se situent en marge des injures, parce qu'elles reposent sur une connivence partagée, un jeu quasiment rhétorique, des rituels établis, des codes. Mais c'est un jeu à risque, et quand elles sont perçues comme blessantes, les vannes peuvent basculer dans l'injure.

4.2. Fonctionnement de l'insulte dans un rapport hiérarchique

Situation dissymétrique

Une remarque d'abord. Il est des formes d'insultes qui finalement n'en sont pas, des jeunes quand ils se parlent entre eux. L'insulte « eh bâtard !» se trouve désémantisée, perd sa valeur illocutoire, pour servir de terme d'adresse, voire d'affection (Lagorgette et Larrivée [10]), ou pour parfois prendre la seule valeur de ponctuant du discours (Caubet [4] Je voudrais saisir ici, dans un jeu de rapports de forces symboliques et dans un cadre dissymétrique, qu'il soit institutionnel ou personnel, le rôle perlocutoire de l'insulte. Dans cette configuration-là, contrairement aux fonctionnements entre pairs, l'insulte arrive généralement en clôture, dans une forme ultime de la montée en tension ; c'est elle souvent, comme nous l'avons observé à travers les rapports d'enseignants (rapports qui relatent les conflits et justifient les punitions prises), qui déclenche la décision de sanction. Elle vient même après toute autre forme d'actes de langage négatifs, comme le harcèlement ou la menace.

Il est à noter que l'insulte use de sa forme essentialisante de façon radicale et réductrice dans les échanges en milieu informel. L'espace de la rue, en ce sens, est caractéristique de la fulgurance de l'insulte. Il s'agit dans de telles situations soit d'exprimer une forme de tension (peur par exemple), soit d'affirmer sa prise de pouvoir et de contrôle sur l'autre dans un effet de négation radicale. C'est ce que l'on observe par exemple dans les altercations entre automobilistes.

Il s'agira ici, à travers un corpus d'interactions en salle de classe de saisir la fonction de l'insulte (pour une analyse interactionnelle détaillée de cet extrait d'un point de vue du rapport de places entre enseignant et élèves, (Auger, Fillol, Lopez, Moïse [2]). Il rend compte d'une transgression affirmée du rapport institutionnel et d'une remise en question de la légitimité de l'enseignant. Il s'agit de montrer comment l'insulte surgit dans une montée en tension et vient, je dirais, en dernier recours d'autres procédés discursifs nombreux qui relèvent davantage de la joute verbale. Nous montrerons donc quels sont les autres procédés utilisés dans la violence verbale, déclencheurs de l'insulte.

Exemplier : extrait de corpus. Enregistrement en salle de classe.F = Fouad. K = Kader. Pr = professeur. C = enquêtrice. ? : élève non identifiable

Les passages marqués en gras sont ceux qui servent davantage l'analyse.

Séquence 1 (mélange de plusieurs voix d'enfants)

F : XX madame

Pr. : regardez bien tous le tableau (*voix)* / pour vendredi

? : bon allez oh (*mélange de la voix du professeur et de celles des élèves*)

F : bon Kader oh
C : c'est comme ça tous les jours / c'est pire que ça d'habitude
? : madame qu'est-ce qu'on écrit
? : madame qu'est-ce qu'on écrit
C : tu tu vas en espagnol toi
? : non en anglais
F : **tout le monde a pas leur carnet / y a que moi** XX
? : madame qu'est-ce qu'on écrit
? : ben pour vendredi
F : même que vous XX
Pr : j'attends que monsieur C. arrive
F : **c'est Mme Brulle que me l'a pris**
? : ça pour vendredi / hein ça pour vendredi // hein
Pr : oui:
? : ça pour vendredi
F: **c'est Mme Brulle dès X elle l'a pris**
(*voix d'élèves)*
F **: oui c'est Mme Brulle**

Séquence 2.
Pr **: mettez tous votre carnet de liaison à côté de vous**
F **: oui c'est Mme Brulle**
K : elle m'a dit Mme Brulle
F : moi XX / Kader /
Pr : mets ton carnet (*mélange de voix)* B. mets ton carnet XXX mets le dessus
? : madame / madame / madame
Pr : voilà / **Kader tais-toi je t'en prie**
K : je montre mon cahier / mon carnet XX tu pas pas
? **:** madame madame
K : madame / pourquoi ils ont pas leur carnet et ils parlent quand même
? : vous allez nous suivre jusqu'à la fin de l'année
C : oui
? : ah ça va être extra ça
C : ça sera pas moi tout le temps y aura une autre personne? : ah oui une jeune là
C : oui
?: X les héros de Batman X
? : pardon madame(*mélange de voix*)
? : madame: / madame:
Pr : XX
K : **tout le monde a pas leur carnet**
F : **allez vous allez me mettre un mot moi aussi / c'est Mme Brulle qui l'a pris**
K : **non non non elle a dit elle le rend lundi à tous**
? : elle a dit pour ceux qui viennent cet après-midi XX chuuuuut
Pr : prenez votre cahier de texte:: / pour écrire le XXXXX depêche-toi Fouad de prendre ton cahier de textes
(*mélange continu de voix d'élèves et de la professeur*)
K : **vous voulez que je parte plus vous donnez un avertissement plus des devoirs plus des rédactions mais: / vous voulez pas de l'or en plus**
? : ça va aller au principal ça / hein madame
C : ça là
? : ou
C : non pas du tout
? : ça va aller où
K: **je vais être renvoyé XX chez Mme Morel**
C : ça va aller pour euh:: (*voix mélangées*) c'est mon travail de X
? : et quand y aura la X vous le mettrez là XXXXXXXXX (*voix*)
K : **tout le monde a pas leur carnet**
? : maman euh madame: ça veut dire quoi n'oubliez pas de tenir compte des conseils X rédaction XX distribués
Pr : je vous ai distribué euh une photocopie sur laquelle je vous ai écrit le barème X
? : oui je vous l'ai rendue moi (*voix*)
Pr : et des conseils prenez une feuille propre prenez-la à l'endroit et pas à l'envers faites des paragraphes
? : ça y est j'ai fini
? : ça y est j'ai marqué
Pr : voilà ce que je veux dire hein

Séquence 3
Pr : **alors je vais vous lire deux rédactions les deux meilleures rédactions que j'ai corrigées**
? : c'est la mienne
Pr : vous allez les écouter (*voix*) alors voilà la première (*voix*) vous les aurez après XXX vous écoutez X hein / alors j'habite aux HLM [É] heureusement il y a mon cousin qui habite à deux minutes de chez moi juste devant l'école primaire et je sors souvent avec lui / je n'ai pas de groupe d'amis je suis surtout avec ma cousine
K : **oh c'est bon on s'en fout**
Pr : parfois après l'école dans le jardin public à côté de chez moi je joue avec des copains et copines de mon ancienne école primaire
K: **parle XX trente sept** *(il joue aux pokemons)*
Pr : **oh: maintenant c'est moi qui vais me charger de vous descendre hein (*toux*) je vais me charger toute seule hein de vous descendre** /
K : **nique ta mère** *(plus bas)*
Pr : **alors taisez-vous tous les deux / c'est bien compris / c'est moi qui vais vous descendre tous les deux (*voix*) K : vous mettez des avertissements**
Pr : **alors taisez-vous / taisez-vous s'il vous plait c'est tout ce que je vous demande / sinon je vous descends moi-même**
? : moi j'ai rien fait madame hein
Pr : **c'est moi qui vous y emmène / et là vous y resterez c'est moi qui vous le dis** /
F : **pourquoi vous dites vous**
Pr: **attention tous les deux**
? : mais moi j'ai rien fait
Pr : **tous les deux parce que vous discutez tous les deux vous me gênez / et vous gênez tout le monde (*brouhaha*) on se tait c'est tout ce qu'on vous demande / écoutez XXXXXXXXX Kader je te descends / euh ou est-ce que j'en étais /**

Séquence 4. (*lecture des rédactions*)

Séquence 5.
Pr : **alors il faudrait m'emmener XX parce que tout à l'heure je l'ai renvoyé**
F : (*voix en arabe*) Kader / t'es renvoyé
Surveillant/éducateur : bon
F : oh mon pote R.
K : X **c'est pas bien ça qu'elle fait hein**
Pr : Fouad
F : quoi chut (*brouhaha*)
F : **oh la la parce qu'on n'a pas le carnet**
Surveillant/éducateur : bon allez (*bruit*) Kader et Fouad allez allez allez
F : et T. (*bruit*) oh la la mais T. il l'a pas aussi
Surveillant/éducateur : Fouad allez
Pr : vas-y X
F. : oh ça m'énerve hein ça
K : **j'y vais pas moi**
Surveillant/éducateur : allez tu prends tes affaires et ton blouson (*bruit*) Fouad Fouad complique pas les choses X allez (*brouhaha*) mais le balance pas en plus? : qu'est-ce qu'il a fait
Surveillant/éducateur: allez / pousse-toi dépêche-toi (*brouhaha*) / allez / dessiner tu peux le faire en étude / allez
K : j'ai pas de cours après je rentre chez moi alors
Surveillant/éducateur : ouais t'as raison allez
? : mais bien sûr (*brouhaha*)
Pr : **vous m'avez fait crier pendant dix minutes alors** Surveillant/éducateur : ouais / bon Fouad oh
? : Mohamed
Surveillant/éducateur : **on répond pas on répond pas on ne répond pas**
Pr : **ne me menace pas Kader**
K : *(adresse en arabe)*
F : arrêtez de crier (*brouhaha*)
Pr : XX ils étaient bien et puis de nouveau XX(*brouhaha*)
C : qu'est-ce qu'il a dit là tout-à-l'heure
? : XX
C: qu'est-ce qu'il a dit
? : **salope**
Surveillant/éducateur : allez
? : madame à quoi ça sert ça
? : ça enregistre tout ce qu'on tout ce qu'on quand on insulte tout ça / ça enregistre(*brouhaha*)
F : je me tiens tranquille / j'étais là-bas
Surveillant/éducateur : pour aujourd'hui c'est trop tard
Pr : **tu m'as interrompue sans arrêt tu m'as fait crier pendant dix minutes / alors non** (*brouhaha*)
F : **dix minutes / je te parlais / elle me dit tu m'as interrompue / c'est quoi ça** (*brouhaha*)

Procédés d'argumentation et d'explication

Nous sommes en cours de français, en 6ème. L'enseignante demande aux élèves de mettre leur carnet de liaison sur la table. Certains élèves ne l'ont pas. Lors de la première séquence enregistrée, un élève, Fouad, interpelle l'enseignante à diverses reprises par l'apostrophe *madame*, pour lui expliquer qu'il n'a pas son carnet. Il fait appel aussi à un sentiment d'injustice et à une demande d'équité *tout le monde a pas leur carnet / y a que moi* [qui vais être puni], signifiant qu'il n'est pas seul dans son cas. Enfin, ses interventions restant sans effet, il tente d'expliquer pourquoi il n'a pas son carnet, *c'est Mme Brulle qui me l'a pris.* Mme Brulle est l'enseignante de mathématiques, qu'ils ont eue l'heure précédente. Nous sommes là dans une tentative de discussion-argumentation sans retour discursif de la part de l'enseignante. En effet, elle ne répond pas. Fouad voudrait expliquer pourquoi il n'a pas son carnet, il est dans une demande de parole, parole qui lui est refusée. Les partenaires s'affrontent autour de l'objet du carnet, à forte valeur symbolique quand on sait l'importance des papiers pour ces élèves. Ne pas avoir son carnet – ses papiers - c'est voir nier son statut d'élève. Fouad et Kader tentent de s'expliquer. Dans la séquence 2, tandis que Kader reprend l'argument explicatif de Fouad, *elle m'a dit Mme Brulle*, l'enseignante s'adresse à lui dans un acte d'autorité, *mets ton carnet / mets ton carnet / mets-le dessus*, puis dans une forme d'oxymore pragmatique (Moïse [16]), *Kader tais-toi je t'en prie.* Kader tente encore de s'expliquer ; il précise la situation, *non non non elle a dit elle le rend lundi à tous.* La tentative explicative ayant échoué, Kader use d'un argument d'équité, *pourquoi ils ont pas leur carnet et ils parlent quand même*, sans plus d'effet. Fouad, de son côté, sollicite au moins à neuf reprises Mme Ravalo sans que celle-ci ne lui adresse directement la parole, alors qu'elle interpelle Kader dès sa première ou deuxième intervention. Effectivement, l'enseignante ressent la demande de Kader après celle de Fouad comme un harcèlement, proche de « l'enfade », harcèlement de la victime obtenu par répétition d'une même séquence verbale ou par accumulation de variantes à partir d'un même acte (Lopez,[12]). Mais ce début d'enfade aurait pu être arrêté par une réponse de l'enseignante. On se rend compte, à partir de l'observation du corpus, combien la demande de discussion de Fouad puis de Kader n'aboutit pas. L'enseignante leur refuse la parole, et en quelque sorte leur statut d'élève. Dans cette impossibilité qu'il a de rejoindre l'enseignante, Kader va alors rejouer un rôle qu'il connaît bien et auquel il se voit, de fait, assigné, celui du jeune de banlieue ; le glissement se produit avec l'intervention sous forme de joute verbale de Kader, *vous voulez que je parte plus vous donnez un avertissement plus des devoirs plus des rédactions mais: / vous voulez pas de l'or en plus*

La joute verbale

L'enseignante a mis Fouad et Kader dans une situation particulièrement difficile avec l'injonction de l'ouverture de la séquence 2, *mettez tous vos carnets à côté de vous*. On peut même se demander si le *tous* ne constitue pas une provocation dans la mesure où la

professeure sait maintenant que certains élèves ne pourront pas respecter cet ordre. Elle place, de fait, Kader et Fouad dans une situation ingérable en tant qu'élève, l'alternative étant : se taire et ne pas obéir à l'injonction réitérée concernant le carnet, avec les conséquences prévisibles ou tenter de se justifier en prenant la parole de force et donc interrompre l'enseignante et gêner le cours, comme elle l'expliquera quand le surveillant interviendra. Il y a là une absence de négociation mais, du coup, des stratégies énonciatives mises en place de la part des élèves.

Donc Kader va réactiver des procédés discursifs qu'il connaît bien, notamment, la « grillade » (Lopez J., [12]), constitutifs (finalement) de fait de cette assignation identitaire implicite. Il va finalement l'utiliser et en utiliser les ressources. Les tentatives de négociation ayant cédé désormais la place à l'affrontement, on peut observer que Kader a une grande maîtrise des formes langagières de la culture des rues, raison pour laquelle il tient maintenant le premier rôle alors que Fouad n'intervient plus que comme comparse. Kader utilise les joutes verbales pour mettre en scène une agressivité croissante. Il ne faut pas oublier que ce type d'interaction vise autant sinon plus le public que l'adversaire. On obtient alors un « baratin » conclu par une « grillade », figure assez rare qui prouve la grande habileté langagière de Kader : *vous voulez que je parte / plus vous donnez un avertissement plus des devoirs plus des rédactions mais: / vous voulez pas de l'or en plus.*

L'indifférence

L'enseignante qui a retrouvé sa place d'autorité suite à l'acte posé, renvoyer Fouad et Kader du cours, va lire, dans la séquence 3, les meilleures rédactions qu'elle a corrigées, *alors je vais vous lire deux rédactions les deux meilleures rédactions que j'ai corrigées*. Kader va intervenir par un *oh c'est bon on s'en fout*, non repris par Mme Ravalo. Effectivement Kader « se fout » non seulement de la lecture des rédactions mais aussi de leur contenu, *alors j'habite aux HLM [É] heureusement il y a mon cousin qui habite à deux minutes de chez moi juste devant l'école primaire et je sors souvent avec lui / je n'ai pas de groupe d'amis je suis surtout avec ma cousine […] / parfois après l'école dans le jardin public à côté de chez moi je joue avec des copains et copines de mon ancienne école primaire*. Cette rédaction montre que son auteur ne fait pas partie du groupe de l'école, donc n'est pas solidaire de Kader ; elle ne veut pas, puisqu'il s'agit d'une fille, se retrouver aux HLM où il y a tous les autres élèves, et donc ceux de sa classe aussi ; elle se rattache à l'école primaire, à ses ami-e-s d'avant. Là encore Kader ne pourra que ressentir une certaine exclusion, exclusion entretenue par l'enseignante qui a choisi cette rédaction comme modèle d'excellence… Kader va se mettre alors à jouer aux Pokémons, acte d'indifférence, mais participant du harcèlement, et qui va exaspérer Mme Ravalo, *oh: maintenant c'est moi qui vais me charger de vous descendre hein (toux) je vais me charger toute seule hein de vous descendre***.** Mme Ravalo réaffirme sa position haute, de sujet détenteur de la parole, *c'est moi qui, je vais me charger toute seule*. Sans vouloir s'attarder, on peut noter la polysémie ici de *descendre*, dans un acte peut-être inconscient alors qu'il s'agit simplement d'amener les élèves dans le bureau du principal qui se trouve à l'étage inférieur. La stratégie identitaire et interactionnelle de Kader et de Fouad trouve ses limites. En effet, l'enseignante, par son insistance sur la personnalisation de l'affrontement, marque bien la domination physique dans sa capacité à emmener, donc à déplacer les deux élèves, *je vais vous descendre*. Elle va reprendre le dessus.

L'insulte

Du point de vue de la mise en scène de Kader (principal maître d'œuvre), le harcèlement, l'enfade et la grillade avaient poussé à bout l'adversaire, ici, l'enseignante. Mais alors que, dans le cadre des échanges entre jeunes, ce type de pratique est relativement sans danger, la cible devant généralement faire face à des enfadeurs protégés par leur supériorité numérique et le soutien du public, les rapports de force au sein d'une classe n'obéissent pas aux mêmes règles. L'enfade va donc porter ses fruits, la victime va être touchée selon un schéma largement éprouvé dans ce type d'interaction, mais ses tourmenteurs ne vont pas pouvoir s'abriter derrière le soutien du groupe. L'enseignante avait réussi un relatif retour à l'ordre et avait commencé à lire des rédactions. Ce qui avait fait jouer l'indifférence à Kader, sans pour autant qu'il s'avoue vaincu.

Kader, refusant la dissymétrie constituante du principe d'autorité, veut inverser les places mais il a désormais usé, sans retour, de toutes les stratégies possibles à sa disposition. Reste alors l'insulte à la fin de la séquence 3, toutefois formulée à voix basse, *nique ta mère*, comme un dernier recours. Cette insulte, en voie de ritualisation, pour certains jeunes entre eux, peut être réactivée à tout moment. Soit quand on veut vraiment, entre pairs, atteindre l'image maternelle de l'interlocuteur, soit, comme c'est le cas ici, quand elle vise à transgresser délibérément l'autorité légitime. L'injure fonctionne avec la force de sa portée sémantique et du tabou sexuel mais elle tend aussi à rabaisser l'enseignante par l'intonation et sans doute aussi parce que l'élève s'engage dans un registre d'insultes qui lui appartient. Elle fait alors semblant de ne pas avoir entendu et use d'ordres et d'interdictions, *taisez-vous tous les deux / taisez-vous / taisez-vous s'il vous plaît c'est tout ce que je vous demande*. À la fin de la séquence 2, Kader a remporté une certaine victoire, et à la fin de la séquence 3, l'enseignante a « sauvé la face », en imposant son principe d'autorité, au-delà de cette première injure donc. Un partout.

Quand l'enseignante revient à la charge de plus belle, il capitule en retournant au registre de la tentative d'explication, *vous mettez des avertissements*. Ayant le sentiment d'avoir remporté une victoire nette sur le terrain de l'affrontement direct et personnel, l'enseignante peut se permettre une normalisation de la situation par le biais d'une dépersonnalisation progressive, *vous me gênez / et vous gênez tout le monde* (brouhaha*) on se tait c'est tout ce qu'on vous demande* / *écoutez* / *Kader je te descends* / *euh où est-ce que j'en étais* /

On peut noter l'extrême violence des échanges et des sentiments qu'ils laissent deviner et la non résolution du conflit. Alors que l'enseignante doit s'imaginer que celui-ci s'est clos dans le retour à la norme de la classe, l'analyse de Kader (et, dans une moindre mesure, de Fouad) doit être toute différente. Dans le cadre de la culture des rues, où il s'est situé par les formes d'opposition qu'il a choisies, la seule issue possible aurait été l'affrontement physique, les atteintes aux faces respectives ayant été trop graves. Le tour de passe-passe opéré par son adversaire et qui a consisté à sembler accepter le terrain des deux jeunes (par le rôle de victime d'abord passive puis recherchant le contact direct) avant d'imposer le retour à la norme institutionnelle dominante n'a pu que remplir l'adolescent de frustration et le laisser en attente d'une opportunité de revanche, d'où la suite, et la deuxième insulte. La deuxième insulte, très personnelle et à connotation sexuelle, *salope* dans sa forme cryptée en arabe, est le dernier stade de l'affrontement verbal avant la violence physique.

Cette insulte marque la résolution du conflit et le dénouement. Le surveillant entre dans la classe. Dans cette séquence, l'apparition d'un nouvel actant, le surveillant, va être l'occasion pour les deux adolescents de tenter de rejouer la partie. Mais le sentiment d'injustice et de frustration accumulées au cours des échanges précédents va conduire Kader à mener le conflit à un terme plus acceptable pour lui : il conclura l'interaction par une insulte en arabe, réaffirmant ainsi non pas son identité, mais une de ses identités, celle qu'on lui a assignée et la seule qui lui permette, dans ces circonstances, de se comporter de manière honorable. Kader « sauve la face », par la recherche de l'affrontement personnel, *ne me menace pas Kader*, par l'insulte finale, dans une forme de théâtralisation - il quittera la scène sur ce mot de clôture –, par une dernière grillade, *j'ai pas de cours après je rentre chez moi alors*. Mais alors qu'une partie avait été remportée par l'enseignante par abandon de ses adversaires, ceux-ci vont, chacun à sa manière, remporter la suivante. Kader, grâce à son insulte, sera sur le terrain le plus proche de la violence physique, Fouad usera de la rhétorique en pointant les incohérences dans l'argumentation de Mme Ravalo, *dix minutes / je te parlais / elle me dit tu m'as interrompue / c'est quoi ça*.

Les valeurs de l'insulte

L'insulte joue de plusieurs effets. Elle est le dernier recours avant l'affrontement physique ou, en tout cas, elle le contient tout en le matérialisant sur un plan symbolique. Kader sait bien que l'effet sera la sanction (sortie de la classe, punition, etc), mais c'est la seule issue interpersonnelle hors de la violence physique.

Par l'insulte, Kader affirme sa prise de pouvoir sur l'enseignante et garde la face auprès du public. Cette prise de pouvoir se joue par la nature même de l'insulte, elle vise avec force à la dévaluation de l'autre. Kader voudrait saisir l'enseignante dans une vérité qui la rabaisserait, façon de se mettre lui-même hors de cause. L'insulte servirait à persuader l'interlocuteur et tout le public, autant que possible, que c'est sa propre nature qui est stigmatisée, et non pas son statut institutionnel. Il vise la personne et non plus l'enseignante, la fonction. L'insulte permet à Kader de prendre position, sans se l'avouer ouvertement, comme la source du jugement évaluatif.

Mais du côté de Kader, l'insulte renforce aussi l'affirmation de soi, comme s'il fallait répondre à l'identité à laquelle on est assigné. Constatons, pour conclure, que ce conflit résulte en grande partie de la représentation stéréotypée que la professeur se fait des deux adolescents et plus particulièrement de Kader, représentation (nous avons analysé aussi la production d'insultes indirectes produites lors d'entretiens par les protagonistes respectifs) qui la conduit à les enfermer dans une identité conçue sous des aspects uniquement négatifs, avec toutes les impasses pédagogiques et communicatives que cela entraîne. L'aspect relatif des positionnements identitaires et des comportements au cours d'une interaction disparaît alors pour laisser la place à une identité absolue et unique, totalement fantasmatique. Pour Kader, c'est une forme de visibilité sociale qui est à l'œuvre, qui traduit le besoin de se construire une identité et de l'afficher : identité refuge dans la cristallisation d'une identité de repli.

5. BIBLIOGRAPHIE

[1] N.Auger, B., Fracchiolla et C.Moïse, ., « Réactions au texte de Michelle Van Hooland » (sous la dir.), Van Hooland, M. (dir.) *Psychosociolinguistique, Les facteurs psychologiques dans les interactions verbales*, L'Harmattan, pages 95-107, 2005

[2] N.Auger, N., V.Fillol, J.Lopez et C. Moïse, 2003 « La violence verbale : enjeux, méthode, éthique ». *France, pays de contacts de langues*, Actes du colloque de Tours, 9 et 10 novembre 2000, 2003

[3] J.Boutet, *Construire le sens*, Neuchâtel, Peter Lang, 1994

[4] D.Caubet, « Du *baba* (papa) à la mère, des emplois parallèles en arabe marocain et dans les parlures jeunes en France », *Cahiers d'études africaines (Langues déliées)*, 163-164, pages 735-748, 2001

[5] J. Chambers, J., *Sociolinguistic Theory*, Oxford, Blackwell, 1994

[6] F. Gadet, « Vers une sociolinguistique des locuteurs », *Le futur de la sociolinguistique européenne*, Sociolinguistica, numéro 14, pages 99-103, 2000

[7] R., Galatalo, M. Mizzau, « Conflit conversationnel et malentendu : quelques relations possibles », *La linguistique* 34-1, pages 151-164, 1998

[8] N.Huston, *Dire et interdire*, Paris, Payot, 2002

[9] C., Kerbrat-Orecchioni, *L'énonciation*, Paris, Armand Colin, 1997

[10] D.Lagorgette et P.Larrivée, « Interprétation des insultes et relations de solidarité », *Langue française*, numéro 144, pages 83-104, 2004

[11] E., Larguèche, *L'injure à fleur de peau*, Paris, L'Harmattan, 1993

[12] J. Lopez, Grillades, enfade et baratin, formes ritualisées de communication chez les jeunes Pailladins, Mémoire de DEA, Université Paul Valéry, Montpellier, 1998

[13] R., Mesthrie. (dir.) *Concise Encyclopedia of Sociolinguistics*, Amsterdam, 2001

[14] C. Moïse, « Pour quelle sociolinguistique urbaine ? », in *Pratiques langagières urbaines, enjeux identitaires, enjeux cognitifs*, VEI Enjeux, numéro 130, Centre de documentation pédagogique, Paris, pages 75-87, septembre 2002

[15] C. Moïse, « Des configurations urbaines à la circulation des langues… ou… les langues peuvent-elles dire la ville ? », *Frontières et territoires urbains, les frontières sociolinguistiques*, Journée internationale de sociolinguistique urbaine, Kénitra, Maroc, 12 décembre 2003, in Thierry Bulot, T. et Messaoudi, L. (dir.), *Sociolinguistique urbaine (frontières et territoires)*, Éditions Modulaires Européennes, Cortil-Wodon, Belgique, *)*, pages 53-80, 2003

[16] C., Moïse, « Postures sociales, violence verbale et difficile médiation », *Les médiations langagières* (Dir. Delamotte-Legrand), Actes du colloque de Rouen, 7-8 décembre 2000, presses de l'université de Rouen, pages 335-349, 2004

[17] H., Ott, « L'approche constructive des conflits », *Cahiers de la réconciliation*, numéro 1-2, 1997

[18] V.,Traverso, *La conversation familière*, Presses universitaires de Lyon, pages 184-193, 1996

[19] M. Van Hooland, *La parole émergente. Approche psychosociolinguistique de la résilience. Parcours théorico-biographique*, L'Harmattan, 2002

[20] G.,Williams, *Sociolinguistics. A sociological Critique*, London, Routledge, 1992

[21] P. Watzlawick, J. Weakland, R. Fisch, *Changements*, Points Seuil, 1975

Session VI

Prosodie

Mardi 13 juin 2006 - 10h00 11h30

Lecture silencieuse et oralisée des phrases relatives : le rôle de la prosodie

Christelle Dodane et Angèle Brunellière***

*Laboratoire Dynamique du Langage, UMR CNRS 5596
14 avenue Berthelot, 69 363 LYON Cedex 07 – France
dodane@isc.cnrs.fr
**Laboratoire de Psychologie Expérimentale, Faculté de Psychologie et des Sciences de l'Éducation de Genève
Bd. Du Pont d'Arve 40, CH-1205 Genève
Angele.Brunelliere@pse.unige.ch

ABSTRACT

The purpose of this study was to determine whether prosody contribute to the integration of the syntactic level during the reading of relative sentences. A behavioural experiment was run on 20 French subjects. First, they had to read sentences silently presented visually (without prosodic markers such as commas). In an additional task using the same procedure, 10 subjects have to read sentences in a loud voice. The comparison between the two tasks reveals that words with the greater reading time in the first study correspond in the second study to a major prosodic phrase boundary. Results suggest that subjects have to restore prosodic contour in order to process syntactically relative sentences. The study of prosody in such a task provides a new methodology to access to the on-line syntactic processing.

1. INTRODUCTION

Pour accéder au sens des phrases relatives telles que *« Le chat qui regarde l'oiseau boit du lait »* (sujet-sujet, désormais SS) et « *Le chat que l'oiseau regarde boit du lait* » (sujet-objet, désormais SO), le sujet doit d'abord attribuer une catégorie grammaticale à chaque mot et ensuite, procéder à une analyse syntaxique qui va lui permettre d'assigner des rôles thématiques aux différents noms et verbes de ces phrases. Si cette assignation est simple à réaliser dans les phrases SS car N1 est l'agent du verbe de la phrase relative (principe de coréférentialité), elle est plus complexe dans les phrases SO car cette fois, N1 est le patient du verbe de la relative. Si la grammaire traditionnelle décrit le fonctionnement de ces phénomènes, elle se fonde entièrement sur une analyse de la langue écrite. Cette focalisation sur l'écrit contribue à masquer le rôle des indices présents dans le signal de parole. En effet, à l'oral, les frontières linguistiques majeures sont marquées par des indices acoustiques et en particulier, prosodiques (paramètres de durée, Fo et amplitude). Ainsi, la fin de la proposition relative correspondant respectivement au mot « *oiseau* » dans la phrase SS et au mot « *regarde* » dans la phrase SO est délimitée à droite par un accent mélodique réalisé par une montée de continuation et un faible allongement final. Le repérage de cette frontière prosodique permet de délimiter la fin de la proposition relative et indique l'arrivée imminente du verbe de la proposition principale. Cette frontière prosodique facilite grandement le traitement de la phrase, mais elle n'apparaît pas à l'écrit. Etant donné l'importance des indices prosodiques dans le traitement de l'information linguistique à l'oral, il est fort probable que le sujet ait besoin de restaurer le niveau prosodique pour pouvoir traiter la langue écrite, notamment au moment de la lecture (passage d'une information de type visuelle aux représentations mentales des formes linguistiques). Or, des études récentes montrent que les signaux prosodiques ne sont pas seulement véhiculés par le signal de parole et par des paramètres acoustiques explicites, mais également par une prosodie de type implicite présente dans le langage en modalité visuelle (Fodor [1-2], Steinhauer & al. [3-4], Quinn & al. [5], Pynte & al. [6], June [7]). Nous posons donc l'hypothèse que lors de la lecture des phrases relatives, les sujets ont besoin de restaurer le contour prosodique pour pouvoir les analyser syntaxiquement et être en mesure d'assigner les différents rôles thématiques. Afin de tester cette hypothèse, nous avons réalisé une étude comportementale comprenant deux tâches, la première en lecture silencieuse où nous mesurons le temps de lecture des différents mots au sein de phrases relatives et la seconde, en lecture oralisée, où nous étudions le contour prosodique des phrases produites par les sujets.

2. MATÉRIEL ET METHODE

2.1. Tâche de lecture silencieuse

Participants

Cette étude a été effectuée sur 20 participants volontaires, 10 femmes et 10 hommes d'un âge compris entre 23 et 33 ans (âge moyen de 26,5 ans +/- 2,52). Tous les participants sont droitiers et présentent une bonne acuité visuelle ou corrigée. Leur langue maternelle est le français et ils n'ont jamais présenté de trouble du langage ou autres troubles neurologiques.

Stimuli

108 phrases ont été construites pour l'expérimentation. Elles se répartissent en trois catégories, 36 phrases relatives SS (du type « *Le chat qui frappe la grenouille regarde l'éléphant* »), 36 phrases relatives SO (du type « *Le chat que la grenouille frappe regarde le singe* ») et 36 phrases de structure syntaxique simple dites « filler » (du type « *Le chien chatouille l'éléphant* ») ou « *Le cheval attrape la tortue et la tortue montre le poisson* »). Les phrases « filler » servent à masquer aux sujets le véritable sujet de l'expérience, c'est-à-dire le traitement

des phrases relatives. Afin que les sujets puissent se concentrer avant tout sur l'analyse syntaxique de ces phrases, l'information de type sémantique a été réduite au maximum. Les phrases relatives sont composées de 9 mots et les phrases « filler » ont une longueur variable, de 5 à 11 mots. Les noms et les verbes ont été choisi en fonction de leur fréquence élevée, mais ils sont de longueurs différentes (mono, bi ou trisyllabiques). Chaque nom et chaque verbe sont présentés dans toutes les positions possibles au sein de chaque type de phrases (SS, SO et filler). L'association des deux verbes présents dans la proposition relative et la proposition principale des phrases relatives SS est la même que dans les phrases relatives SO.

Procédure expérimentale

Les sujets sont assis à 50 cm d'un écran d'ordinateur sur lequel apparaissent les phrases. Avant chaque mot, un point de fixation apparaît au centre de l'écran. Une fois que le sujet a lu le premier mot, il appuie sur une touche du clavier afin de provoquer l'apparition du mot suivant. Un tiers des phrases est suivi d'une question portant sur l'assignation thématique des rôles. Le sujet doit alors déterminer si la question relate les mêmes rôles thématiques que la phrase précédente. Les questions sont réparties de manière équivalente pour les différents types de phrases. Cette procédure expérimentale a été réalisée avec le logiciel E-Prime.

Analyses statistiques

Deux analyses de variance (ANOVA à mesures répétées) ont été réalisées dans le but de déterminer l'influence de la complexité des phrases (1^er facteur SS-SO) et la position des différents mots (2^ème facteur : position des verbes – position des noms) sur le temps de lecture (exprimé en ms). Par ailleurs, pour chaque facteur, des analyses Post-Hoc de type LSD ont été réalisées.

2.2. Tâche de lecture à haute voix

Participants

Une tâche supplémentaire de lecture à haute voix est demandée à la moitié des sujets ayant participé à l'étude précédente (5 femmes et 5 hommes, d'un âge moyen de 26,8 ans +/-2,86).

Stimuli

Pour cette tâche, 18 nouvelles phrases sont construites avec les mêmes caractéristiques que lors de la tâche précédente. Elles se répartissent également en trois catégories, 6 phrases SS, 6 phrases SO et 6 phrases « filler ».

Procédure expérimentale

La procédure expérimentale utilisée est similaire à celle de la tâche de lecture silencieuse à la différence près que les dix sujets doivent lire les phrases à haute voix et qu'ils sont enregistrés.

Enregistrements et analyses acoustiques

Chaque sujet a été enregistré durant la totalité de la tâche de lecture et ses productions ont été directement transférées sur le disque dur de l'ordinateur via un microphone unidirectionnel Philips SBC-MD695. Les phrases ont été échantillonnées à 44KHz, 16 bits en mono. Les analyses acoustiques ont été réalisées avec le logiciel Praat et ont consisté à relever la valeur de Fo (en Hz) à la fin de la dernière syllabe de chaque mot et la durée (en ms) des mots, syllabes et pauses à partir d'une segmentation préalable du signal de parole. Afin de normaliser les données pour les 10 locuteurs étudiés, nous avons converti les valeurs de Fo en demi-tons en utilisant la formule de conversion S=(log(X)-log(50)/(log(2)/24), X étant la valeur en Hz à convertir et 50, la fréquence de référence. Nous avons également déterminé la durée relative de chaque syllabe en calculant le rapport entre sa durée et celle de la syllabe la plus longue de la phrase (d'un rapport égal à 1).

3. RÉSULTATS

3.1. Tâche de lecture silencieuse (données comportementales)

L'analyse statistique portant sur les noms révèle un effet de la complexité ($F(1,19)=9,276$, $p<0,05$), un effet de la position ($F(2,38)=6,678$, $p<0,05$) et un effet de l'interaction complexité/position ($F(2,38)=6,483$, $p<0,05$). Comme le montre la figure 1, dans les phrases SS, le nom au sein de la proposition relative présente un temps de lecture plus élevé que l'ensemble des autres noms ($p<0,05$), ce qui correspond à un surcroît de traitement au niveau cognitif lié à l'intégration de l'information syntaxique (fin du groupe, attente de la proposition principale). Il se trouve que ce mot est également localisé à la frontière d'une unité intonationnelle (IU), frontière qui est prosodiquement très marquée à l'oral. Elle est en effet délimitée à droite par un accent mélodique ou « pitch accent » réalisé par une montée de continuation et un faible allongement final.

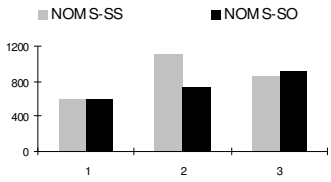


Figure 1 : Temps de lecture en ms en fonction de la position des noms (1 : nom précédant le pronom relatif ; 2 : nom dans la relative ; 3 : nom clôturant la phrase).

Il apparaît également que le nom précédant le pronom relatif et le nom clôturant la phrase possèdent des temps de lecture identiques dans les phrases SS et SO ($p>0,05$). Par ailleurs, on peut remarquer que quel que soit le type de phrases, le temps de lecture des noms précédant le

pronom relatif est plus faible que celui clôturant la phrase ($p<0,05$). Ce résultat était attendu étant donné que le dernier mot de la phrase implique un processus cognitif d'intégration de l'ensemble de la phrase. Dans le cas des verbes (voir figure n°2), nous observons un effet de la complexité ($F(1,19)=12,528$, $p<0,001$), une absence d'effet de la position ($F(1,19)=0,21$, $p>0,05$) et un effet de l'interaction complexité/position ($F(1,19)= 6,267$, $p<0,05$).

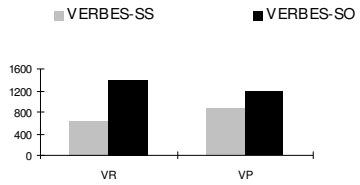


Figure 2 : Temps de lecture (TL) en ms en fonction de la position des verbes (VR : verbe relatif, VP : verbe principal).

Le temps de lecture des deux verbes présents au sein des phrases SO est supérieur à celui des verbes au sein des phrases SS ($p<0,05$). Dans les phrases relatives SO, la frontière prosodique est justement localisée à la fin du verbe de la proposition relative à l'oral (matérialisée là aussi par une montée mélodique et un faible allongement final). L'augmentation du temps de lecture sur le verbe de la relative s'explique par un surplus de traitement cognitif. En outre, le traitement semble se poursuivre sur le verbe de la principale, qui suit immédiatement le verbe de la relative dans les phrases SO.

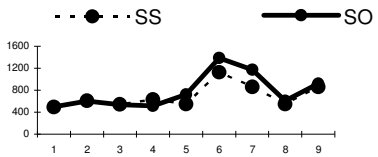


Figure 3 : Temps de lecture (TL) en ms en fonction de la position des mots (numérotés de 1 à 9 par ordre d'apparition au sein de la phrase).

3.2. Tâche de lecture oralisée

Sur la figure n°4, nous avons reporté les contours joignant les différentes cibles mélodiques associées à la syllabe finale de chacun des 9 mots des phrases SS et des phrases SO, exprimées en demi-tons. Cette méthode fournit un contour mélodique lacunaire et lissé de façon grossière, mais elle renseigne sur les cibles mélodiques atteintes successivement et fournit en particulier le point possédant la hauteur la plus élevée pour chaque type de phrase. Nous appellerons ce point le centre intonatif de la phrase. Il se trouve qu'il est localisé pour les phrases SS, sur la fin du nom clôturant la proposition relative et pour les phrases SO, sur la fin du verbe de la proposition relative ($6^{ème}$ position dans les deux types de phrases). Le contour obtenu présente donc dans sa deuxième moitié un profil similaire à celui de la figure n°3 pour les temps de lecture. Ainsi, on peut remarquer que le centre intonatif correspond au temps de lecture le plus élevé, et ce, quelque soit le type de phrases.

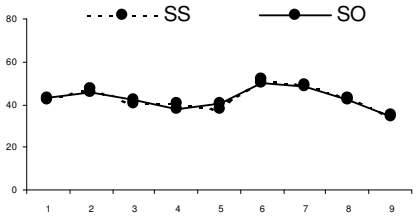


Figure 4 : Contour joignant les différentes cibles mélodiques associées à la syllabe finale de chacun des 9 mots des phrases SS et des phrases SO, exprimé en demi-tons.

En raison du caractère non écologique de la tâche de lecture (présentation des mots les uns derrière les autres), la durée moyenne des syllabes est particulièrement élevée (385 ms pour les phrases SS et 374 ms pour les phrases SO). Cependant, il existe des différences liées à la position relative des différents mots au sein des phrases. Pour les phrases SS (figure n°5), la dernière syllabe du verbe de la relative est particulièrement allongée ($4^{ème}$ position : 520 ms), ce qui coïncide sûrement avec un moment d'intégration de l'information syntaxique du début de phrase (traitement de l'information portée par le pronom relatif par exemple). Le verbe de la principale est également très allongé ($7^{ème}$ position : 503 ms). Il se trouve tout de suite après le centre intonatif de la phrase et nous avons vu précédemment que cette position correspond au $2^{ème}$ temps de lecture le plus élevé (figure n°3).

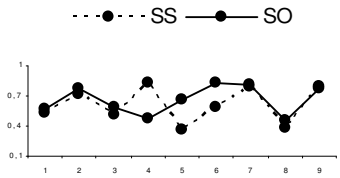


Figure 5 : Durée relative de la dernière syllabe de chacun des 9 mots des phrases SS et des phrases SO.

Pour les phrases SO, la dernière syllabe du verbe de la principale est très allongée (451 ms), ce qui correspond au centre intonatif et au temps de lecture le plus élevé. Par contre, la dernière syllabe du verbe de la relative est également très allongée ($6^{ème}$ position : 472 ms), ce qui correspond là aussi à un temps de lecture élevé (figure n°3). Il semble donc que le sujet doive marquer un temps d'arrêt sur ces deux verbes successifs, afin d'intégrer l'information syntaxique de la phrase SO. Sur le tracé de la figure n°6, on peut suivre la durée (en ms) des pauses après chaque mot successif.

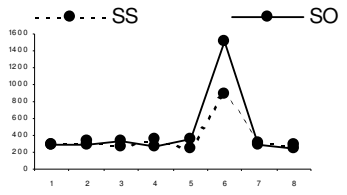


Figure 6 : Durée des pauses (exprimée en ms) entre chacun des 9 mots des phrases SS et des phrases SO.

Ainsi, on relève une pause très élevée après le centre intonatif des phrases SS (878 ms) et SO (1522 ms). Cette position semble être le lieu privilégié de l'intégration syntaxique car tous les indices étudiés voient leur valeur atteindre leurs maxima à son approche. En outre, la durée de la pause est beaucoup plus élevée pour les phrases SO que les phrases SS, ce qui traduit un traitement syntaxique plus complexe pour les phrases SO. Cette information converge avec l'effet de complexité relevé pour les temps de lecture.

4. DISCUSSION

On aurait pu supposer que le caractère artificiel de la tâche de lecture allait totalement perturber le contour prosodique « naturel » des phrases, voire que ce contour serait absent. Or, tout se passe comme si le sujet avait besoin de faire appel au contour prosodique pour traiter l'information syntaxique des phrases relatives SS et SO et de cette façon, être capable d'assigner les différents rôles thématiques au sein des phrases. Ces résultats vont dans le sens de l'hypothèse de la prosodie implicite (Fodor [1-2]), selon laquelle les lecteurs imposeraient un contour prosodique au texte qu'ils lisent silencieusement. La prosodie serait ainsi traitée comme faisant partie intégrante du signal d'entrée et pourrait donc influencer la résolution de l'ambiguïté syntaxique d'une façon similaire à l'oral. Par ailleurs, le traitement des frontières prosodiques à l'oral et celui de la ponctuation à l'écrit suscitent le même type de réponse précoce du cerveau, ce qui suggèrerait un mécanisme commun dans les deux types de traitement (Steinhauer [4]). Selon Fodor [1], la prosodie implicite fait partie intégrante de l'écrit et peut influencer la résolution de l'ambiguïté prosodique de la même façon que la prosodie présente de façon explicite dans le signal de parole. Ainsi, l'attachement des propositions relatives diffère en fonction des langues lorsque leur tête est un syntagme nominal complexe (N1 de N2 - par exemple « *la fille des français* » dans la phrase « *La fille des français qui entre/entrent* »). Fodor [1] explique ces divergences en fonction des différences prosodiques entre les langues. En effet, la prosodie projetée par défaut sur les phrases est déterminée par les principes phonologiques spécifiques à chaque langue. En cas d'ambiguïté syntaxique, la préférence est accordée à l'analyse syntaxique associée au contour prosodique implicite le plus naturel de la langue. Cette hypothèse semble confirmée expérimentalement par des études en lecture silencieuse (Quinn & al. [5], Pynte & al. [6], June [7]) qui révèlent des différences entre les langues dans le traitement syntaxique des phases relatives, mais surtout, une interaction très forte entre la structuration prosodique et l'attachement des propositions relatives. Le rôle de la prosodie semble à ce point important dans l'analyse syntaxique que le cerveau a besoin de la générer lorsqu'elle est absente du signal en modalité auditive (Herrman & al. [8], Meyer & al. [9]). Ce mécanisme s'apparente à un phénomène de type Gestalt où le sujet génère automatiquement les informations prosodiques absentes du signal entrant.

5. CONCLUSION

Nos résultats convergent avec ceux de la littérature récente pour montrer l'importance de la prosodie dans les tâches de lecture aussi bien silencieuses que oralisées. Mais si la prosodie semble bien intervenir lors de l'analyse syntaxique, les avis divergent quant à son rôle exact. Ainsi, il reste encore à déterminer si elle intervient à un niveau précoce pour permettre l'analyse syntaxique ou si le traitement prosodique intervient en parallèle au traitement syntaxique et influence la résolution de l'ambiguïté structurelle.

BIBLIOGRAPHIE

[1] J. Fodor. Psycholinguistics cannot escape prosody. In Proc. Intl. Conf. on Speech Prosody, pages 83-88, 2002.

[2] J. Fodor. Learning to parse. *Journal of Psycholinguistic Research*, 27:285-318, 1998.

[3] K. Steinhauer, K. Alter and A. Friederici. Prosodic boundaries, comma rules and brain responses: The Closure Positive Shift in the ERPs as a universal marker for prosodic phrasting in listeners and readers. *Journal of Psycholinguistic Research*, 30: 267-295, 2001.

[4] K. Steinhauer. Electrophysiological correlates of prosody and punctuation. *Brain and Language*, 86:142-164, 2003.

[5] D. Quinn, E. Fernandez, R. de Almeida, S. Bradley and J. Fodor. Prosodic phrasing predicts RC attachment in French and English silent reading In Proc Amlap Conf, pages, Saarbrucken, 2001.

[6] J. Pynte, J. and S. Colonna. Decoupling syntactic parsing from visual inspection: The case of relative clause attachment in French. In Reading as a Perceptual Process, A. Kennedy (eds), 529-547, 2000.

[7] S. June. Prosodic phrasing and attachment preferences. *Journal of Psycholinguistics Research*, 32, pages 219-249, 2003.

[8] C. Herrman, A. Friederici, U. Oertel, B. Maess, A. Hahne and K. Alter. The brain generates its own sentence melody: A Gestalt phenomenon in speech perception. *Brain and Language*, 85:396-401, 2003.

[9] M. Meyer, K. Steinhauer, K. Alter, A. Friederici and Y. von Cramon. Brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain and Language*, 89:277-289, 2004.

La courbe de F_0 des sonantes initiales de syllabe joue-t-elle un rôle prosodique ? Etude-pilote de données d'anglais britannique

Alexis Michaud, Barbara Kühnert

Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS/ Paris 3 Sorbonne Nouvelle
et Institut du Monde Anglophone, Paris 3 Sorbonne Nouvelle
alexis.michaud@univ-paris3.fr, barbara.kuhnert@wanadoo.fr

ABSTRACT

Several recent publications raise the issue whether the F_0 curve of syllable-initial sonorants can play a prosodic role. The experimental evidence adduced in the present pilot study consists of 15 C_1VC_2 words, where C_1 = /p/, /b/ or /m/, V = /ɑ:/, /i:/, /u:/, and C_2 = /t/; these words were said twice inside a carrier sentence by four speakers of Standard Southern British English. Comparison of the F_0 curves of the /m/-initial syllables with those of the obstruent-initial syllables suggests that only the part of the F_0 curve which corresponds to the syllable *rhyme* is to be taken into account at the stage of the interpretation of the word's prosodic information.

1. INTRODUCTION

1.1. La question générale du rôle prosodique de la courbe de F_0 des consonnes ; le cas des sonantes initiales de syllabe

Il est généralement admis que les consonnes occlusives voisées ne sauraient être porteuses d'une mélodie contrôlée par l'énonciateur. Au plan phonétique, leur voisement est difficile à maintenir ; *a fortiori*, le contrôle de la fréquence fondamentale (ci-après F_0) paraît très difficile pendant ces consonnes. En revanche, il est possible de réaliser des modulations de F_0 pendant l'articulation d'une sonante telle que le [m] de « ma » [ma]. Pour autant, la portion de courbe de F_0 portée par cette consonne peut-elle être porteuse d'information linguistique ? Cette question a un enjeu pour les études prosodiques : ainsi, une syllabe [ma] dont la F_0 s'élève au cours du [m] et descend au cours du [a] peut, selon que l'on exclut ou non la portion de F_0 correspondant à l'initiale, être considérée comme portant un schéma mélodique montant-descendant, ou une simple descente.

La discussion sera limitée à une comparaison entre l'*accent lexical* anglais et le *ton lexical* de certaines langues asiatiques, tous deux lexicalement associés à une syllabe.

1.2. La situation dans des langues à tons lexicaux : la syllabe est divisée en initiale *et* rime *; le ton est porté par la rime*

Dans l'étude des langues à tons d'Asie du Sud-Est, dont la structure syllabique est relativement simple, la syllabe est généralement divisée en *initiale* et *rime*. Dans une syllabe /man/, l'initiale est /m/ et la rime /an/ (pour plus de détails voir Sagart [14, p. 35], et références citées). Le ton appartient à la syllabe, mais est porté par la rime. Les descriptions de nombreuses langues font écho à cette division en initiale et rime (au sujet du yorùbá, langue de la famille niger-congo, voir Laniran [10, p. 61] ; au sujet du danois : Gårding [2, p. 137]).

1.3. L'approche en termes de courbes continues de F_0 : prise en compte de la syllabe dans son intégralité

Plusieurs études récentes, portant sur des langues variées, considèrent que la courbe de F_0 portée par les consonnes initiales sonantes est partie intégrante de la courbe de F_0 qui caractérise la syllabe dans son entier (voir Ladd [9] et références citées, Xu [15]). Ces études reposent sur l'idée qu'à un certain niveau d'abstraction, la hauteur serait une ligne continue, en dépit du non-voisement de certains segments. Sur cette base, des consonnes continues voisées sont préférentiellement employées dans les expériences. Celles-ci se concentrent essentiellement sur l'étude de différences fines dans *l'alignement temporel des courbes avec les segments* (par exemple Kohler [8]). Ces études n'ouvrent pas nécessairement le débat avec l'idée exposée en section 1.2 ; or ces deux conceptions sont en un sens contradictoires. Ainsi, une étude récente avance l'idée selon laquelle, en mandarin, l'alignement entre courbes de F_0 et segments aurait lieu par rapport au début de la syllabe (Xu [16, p. 321]), ce qui est en contradiction avec plusieurs travaux expérimentaux sur cette même langue (dont Hallé *et al.* [3] et Howie [7]).

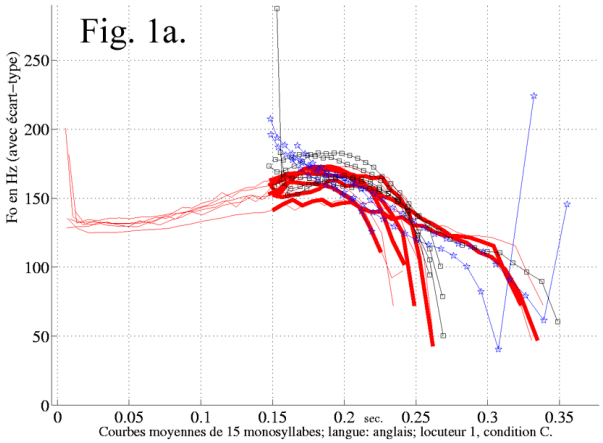
L'expérience-pilote proposée ici se fonde sur des données anglaises. Elle consiste à comparer les courbes de F_0 de deux ensembles de monosyllabes, les uns à initiale sonante, les autres à initiale occlusive.

2. MÉTHODE

2.1. Corpus et locuteurs

Quatre locuteurs d'anglais britannique (ci-après M1 à M4) ont été enregistrés au laboratoire de phonétique de l'Université de Cambridge. Tous quatre étaient étudiants de Licence en linguistique. Ils ont été rétribués pour leur participation. Les données utilisées ici constituent un sous-ensemble du corpus qu'ils ont enregistré (voir Michaud [12]) : au stade de la présente étude-pilote, l'attention se concentre sur la différence entre trois consonnes initiales de même ordre et de série différente, /p/, /b/ et /m/. Ont été seuls retenus neuf monosyllabes composés d'une consonne initiale bilabiale (/m/, occlusive *lenis* /b/, non voisée dans ce contexte, ou occlusive *fortis* /p/, phonétiquement aspirée), d'une voyelle /ɑ:/, /i:/, /u:/, et d'une occlusive finale /t/. Les mots étaient : (1) *mart,* (2) *meet,* (3) *moot ;* (4) *part,* (5) *Pete,* (6) *poop* (en remplacement de **poot*) *;* (7) *bart,* (8) *beet,* (9) *boot.* Certains apparaissent deux fois dans la liste présentée aux locuteurs, ce qui donne en tout 15 items : 6 à initiale /m/, 5 à initiale /b/, 4 à initiale /p/. Ces mots ont été lus deux fois dans la même phrase-cadre avec des indications de contexte différentes (choisies pour une comparaison entre trois langues ; *ibid.* [12]). Dans ce qui suit, la tâche 1 sera désignée comme condition S (« soignée ») et la tâche 2 comme condition I (« insistante »), étiquettes qui n'ont pas valeur définitoire.

Task 1: You're teaching a foreign student who made a mistake when reading a word. (Class context.) Read each item inside the carrier sentence, making a long pause (breathing in and out once) in-between sentences: *Look, this is ___ here.*



Les courbes des syllabes comportant une syllabe initiale /p/ (étoiles) ont une allure globalement descendante. Plusieurs des courbes des syllabes à initiale /b/ (carrés) comportent une courte montée en début de voisement. Les courbes correspondant à

Task 2: A child who is learning to read has asked you how to pronounce this word time and again; (s)he asks you yet another time; you answer, less patiently: *Look, this is ___ here!* Remember to make a long pause (breathing in and out once) in-between sentences.

Afin d'obtenir une mesure très précise de F_0 (ainsi que des indices sur la qualité de voix, non utilisés ici), un enregistrement électroglottographique a été réalisé simultanément avec l'enregistrement audio.

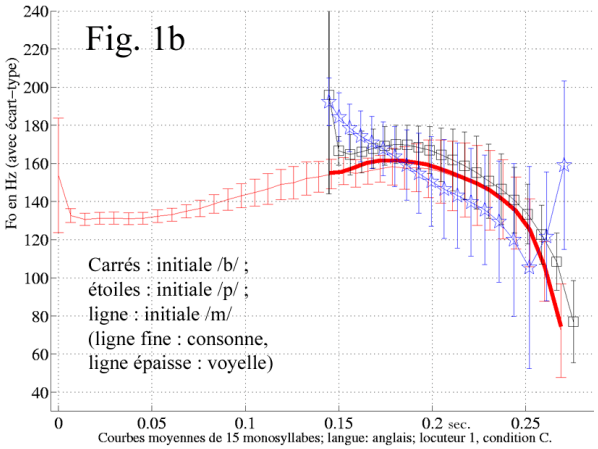
2.2. Analyse des données

Les bornes de début et de fin de chaque syllabe, et les frontières entre consonnes initiales nasales et voyelles, ont été déterminées sur la base de l'inspection et de l'écoute du signal audio, et avec l'aide d'un spectrogramme dans certains cas. La F_0 a été calculée par la détection des pics positifs (« pics de fermeture glottique ») sur la dérivée du signal électroglottographique (au sujet de cette méthode, voir Henrich *et al.* [5, 6] et références citées) ; au sujet des programmes créés pour l'analyse du signal électroglottographique, et pour le calcul de courbes moyennes, voir Michaud [11] et http://voiceresearch.free.fr/egg/.

3. RÉSULTATS ET DISCUSSION

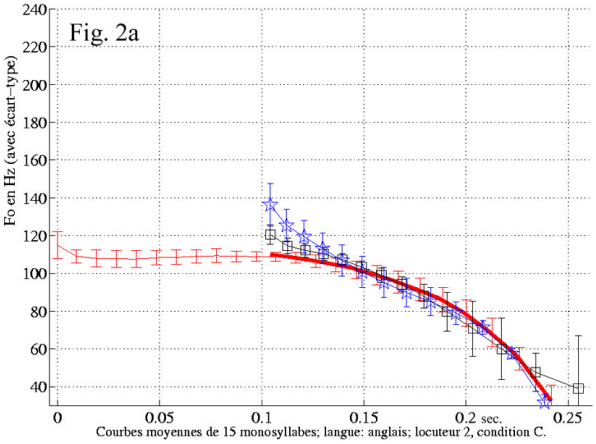
3.1. Commentaire des figures

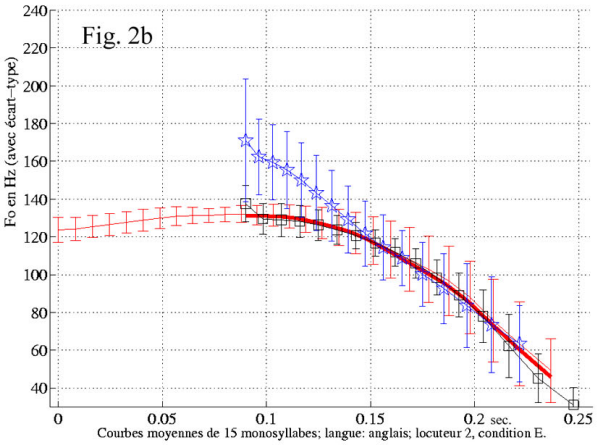
La figure 1 montre, pour un locuteur (M1) et dans une même condition de lecture (condition S), les courbes des trois types de monosyllabes : courbes brutes sur la figure 1a, courbes moyennes sur la figure 1b. Chaque point correspond à un cycle glottique.



l'intégralité des syllabes incluant une initiale /m/ (ligne simple) présentent une allure différente de celle des deux ensembles précédents : une partie légèrement montante (ou parfois simplement égale) précède la descente.

La figure 2 met en regard, pour le locuteur 2 et pour chacune des deux conditions de lecture, quatre courbes : la courbe moyenne obtenue sur les syllabes à initiale *lenis* /b/ ; celle correspondant à l'initiale *fortis* /p/ ; celle obtenue sur les syllabes à initiale continue voisée /m/, en incluant la totalité de la syllabe (consonne initiale plus voyelle) ; et enfin la courbe obtenue sur ces mêmes syllabes à initiale /m/ lorsque seule la partie vocalique est prise en compte. Afin d'évaluer le degré de proximité entre les courbes sur la voyelle des syllabes à /m/ initial et sur la voyelle des syllabes à /b/ ou /p/ initial, les rimes de ces trois types de syllabes ont été alignées par leur point d'origine.





Faute de place, il n'est pas possible de présenter ici les figures représentant les données de chacun des quatre locuteurs ; ces figures sont réunies sur une planche en couleurs disponible à l'adresse suivante : http://ed268.univ-paris3.fr/lpp/pages/EQUIPE/michaud/JEP2006/MichaudKuehnertJEP06_figures.pdf

3.2. Similitude des courbes sur les rimes

Le contexte d'énonciation étant le même pour tous les items, leur courbe de F_0 représente, à un certain niveau d'abstraction, un même phénomène linguistique, le même type de réalisation d'un accent lexical initial de mot (ci-après désigné comme A). Les différences d'allure et de longueur entre courbes de F_0 des syllabes à initiale /p/, /b/ d'une part, à initiale /m/ d'autre part, peuvent s'interpréter d'au moins deux façons. (1) La courbe des syllabes à initiale sonante fournirait l'image la plus complète de la réalisation de A, tandis que la courbe de F_0 des syllabes à initiale non voisée serait incomplète, la courbe continue « sous-jacente » ne pouvant se manifester phonétiquement pendant la consonne du fait des propriétés de la consonne (obstruente et non voisée). (2) La courbe des syllabes à initiale obstruente fournirait une image complète de la réalisation de A ; la courbe des syllabes à initiale /m/ devrait être stylisée en ne retenant que la partie correspondant à la rime, afin de faire ressortir la portion de la courbe de F_0 qui correspond à la réalisation de A. La courbe de F_0 pendant la consonne initiale reflèterait uniquement la préparation de la réalisation de A (en l'occurrence, un *mouvement vers le point de départ d'une courbe descendante*).

En suivant l'hypothèse (1), qui nous paraît correspondre à l'orientation évoquée en section 1.2, on est amené à décrire la réalisation de A comme comportant deux parties : une partie stable (légèrement montante) suivie d'une descente. Cette conclusion ne s'accorde pas avec le cadre du modèle britannique d'études intonatives, dans lequel la réalisation attendue dans ce contexte est une descente *(fall)*. En revanche, suivant l'hypothèse (2), inspirée par la division entre initiale et rime, la réalisation de A est une descente, dans tous les cas ; une fois la courbe des syllabes à /m/ initial réduite à la portion de courbe correspondant à la rime, une certaine similarité (de longueur et d'allure de la courbe) ressort entre les trois sous-ensembles de courbes. La forte différence de longueur entre les courbes avec et sans consonne initiale correspond globalement à la durée du /m/ initial. Les observations visuelles réalisées sont donc compatibles avec l'idée traditionnelle selon laquelle la F_0 au cours des initiales voisées ne joue pas de rôle linguistique, la partie de la courbe coïncidant avec les segments de la rime constituant à elle seule l'unité porteuse des phénomènes prosodiques (qu'il s'agisse, selon les langues, d'un ton, d'un accent, ou de phénomènes intonatifs).

3.3. Hypothèses au plan perceptif

Des recherches récentes montrent que les auditeurs peuvent faire usage d'indices phonétiques ténus (Hawkins [4]), ce qui peut rendre suspect le choix de ne retenir, à un certain niveau d'analyse, que la portion de courbe de F_0 portée par la rime. Néanmoins, il est connu qu'un indice perceptif peut être utilisé au niveau segmental et être en revanche écarté (par *compensation perceptive*) au niveau prosodique, une même information étant utilisée différemment en fonction de la tâche concernée. Ainsi, Reinholt Petersen [13] montre que les perturbations locales de la courbe de

fréquence fondamentale dues à l'influence des consonnes peuvent être utilisées par les auditeurs pour l'identification des consonnes, tandis que ces mêmes auditeurs en feraient abstraction dans leur perception de la prosodie.

En outre, l'idée selon laquelle la courbe de F_0 portée par l'initiale de syllabe ne jouerait pas de rôle prosodique ne revient nullement à nier le rôle intonatif que peut jouer la réalisation des consonnes initiales de syllabe, rôle notamment mis en lumière par Fónagy [1, pp. 88-106] : la longueur d'une consonne, et le détail de son articulation aux niveaux sous-glottique, glottique et supraglottique, peut être porteur d'informations d'ordre attitudinel/émotionnel, et peut également être l'un des indices du découpage de l'énoncé en constituants. Il paraît en outre plausible que les auditeurs puissent utiliser la courbe de F_0 de sonantes initiales comme indice mélodique secondaire.

4. CONCLUSION ET PERSPECTIVES

L'expérience-pilote nous conduit à conclure que dans le cas des monosyllabes anglais étudiés, la portion de la courbe de F_0 portée par la consonne initiale /m/ doit être écartée pour que ressorte, dans son unité, le phénomène prosodique qui se réalise sur la syllabe accentuée *(nucleus)*, en l'occurrence une descente *(fall)*.

La poursuite du travail consistera à se fonder sur des données articulatoires pour affiner l'étude de la transition entre consonne et voyelle. Il paraît nécessaire de recourir à plusieurs méthodes exploratoires qui se complètent, correspondant au plan des phénomènes acoustiques, supraglottiques et glottiques.

Il apparaît en outre indispensable de mettre en place un protocole expérimental comprenant des tâches suffisamment variées pour qu'il soit possible de généraliser au sujet de l'effet de la composition phonémique de la syllabe, l'objectif étant de démêler l'action des universaux phonétiques (facteurs aérodynamiques et physiologiques), d'une part, et celle du système linguistique, d'autre part.

5. REMERCIEMENTS

Vifs remerciements à Francis Nolan pour son accueil à Cambridge, à Geoffrey Potter pour son aide technique, et aux relecteurs de cette communication.

BIBLIOGRAPHIE

[1] I. Fónagy. *La Vive voix : essais de psycho-phonétique*. Payot, Paris, 1983.

[2] E. Gårding. Intonation in Swedish. *Intonation Systems: A Survey of Twenty Languages*, D. Hirst et A. Di Cristo (éds.), Cambridge University Press, Cambridge, 1998, pp. 112-130.

[3] P. Hallé. Evidence for tone-specific activity of the sternohyoid muscle in Modern Standard Chinese. *Language and Speech*, 37:103-124, 1994.

[4] S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *J. of Phonetics*, 31:373-405, 2003.

[5] N. Henrich, C. d'Alessandro, M. Castellengo, et B. Doval. On the use of the derivative of electroglottographic signals for characterization of non-pathological voice phonation. *J. of the Acoust. Soc. of America* 115(3):1321-1332, 2004.

[6] N. Henrich, C. Gendrot, G. Schade, F. Muller et R. Expert. Characterization of features observed on the derivative of EGG signal by the use of high speed cinematography. *International Conference on Voice Physiology and Biomechanics*, Marseille, 18-20 août 2004.

[7] J. M. Howie. On the domain of tone in Mandarin. *Phonetica,* 30:129-148, 1974.

[8] K. J. Kohler. Terminal intonation patterns in single accent utterances of German: phonetics, phonology, and semantics. In *Studies in German Intonation*, AIPUK n°25, Kiel (Allemagne), 1991.

[9] R. Ladd, D. Faulkner, H. Faulkner, et A. Schepman. Constant "segmental anchoring" of F_0 movements under change in speech rate. *J. of the Acoust. Soc. of America*, 106:1543-1554, 1999.

[10] Y. O. Laniran. *Intonation in Tone Languages: the Phonetic Implementation of Tones in Yorùbá*. Ph. D., Cornell University, 1992.

[11] A. Michaud. Final consonants and glottalization: new perspectives from Hanoi Vietnamese. *Phonetica*, 61(2-3):119-146, 2004.

[12] A. Michaud. *Prosodie de langues à tons (naxi et vietnamien), prosodie de l'anglais : éclairages croisés*, thèse de doctorat, Univ. Paris 3, 2005.

[13] N. Reinholt Petersen. Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica*, 43:31-42, 1986.

[14] L. Sagart. *Les dialectes gan. Etudes sur la phonologie et le lexique d'un groupe de dialectes chinois*. Langages croisés, Paris, 1993.

[15] Y. Xu. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, 55:179-203, 1998.

[16] Y. Xu et E. Q. Wang. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*. 33:319-337, 2001.

Le focus prosodique n'est pas que déictique : le modèle VID (Valence-Intensité-Domaine)

Aubergé, V. & Rilliard, A.

Institut de la Communication Parlée UMR 5009 CNRS-INPG- Université Stendhal, Grenoble, France
{auberge, rilliard}@icp.inpg.fr

ABSTRACT

This paper summarizes several perception experiments showing that the morphology of the prosodic focus conveys more information than the only deictic information: (1) the binary valence - yes/no focus – which is perceptively quite categorical (a magnet effect is clear on the basis of an identification and a discrimination experiment [1]), (2) the intensity information, used by the speaker to give his preference for one of two focused elements, (3) the information of the focus domain, that are some segmentation cues about the focused element (phonological unit or word unit), which are perceptively identified by listeners. The morphological cues revealing Valence-Intensity-Domain are observed in particular in morphing procedure making clear the thresholds of quite-categorical behaviors.

1. INTRODUCTION

La focalisation dans la chaîne verbale peut être définie sommairement comme une fonction utilisant des matériaux communicatifs dont la prosodie et la syntaxe pour faire émerger une information nouvelle ou préciser un élément mis en contraste. Ainsi, Rossi [9] considère la focalisation comme un ensemble d'outils qui permettent au locuteur de hiérarchiser l'information et de faire émerger un élément spécifique sur lequel le locuteur veut attirer l'attention de l'auditeur. La fonction d'emphase et de focus lexical sont présentées comme deux fonctions exclusive dans la littérature [5]. La fonction d'emphase est surtout reliée à des traits expressifs comme le degré d'intérêt et est décrite plutôt comme une fonction gradiente, tandis que la fonction de focalisation est binaire (présence/absence de focus). Le focus prosodique est très robuste : un focus lexical est réalisé en français (et peut-être universellement [3]) par une proéminence acoustique, principalement tonale, sur la première syllabe du mot (sauf dans des cas particuliers de conflits ou d'effet stylistique). Pourtant un mot contenant un tel ton sur n'importe laquelle de ses syllabes, est néanmoins perçu comme focalisé [2]. Jackendoff [4] a montré que le domaine du focus peut être méta-linguistique quand il renvoit à des valeurs communicatives différentes du focus lexical : l'« ordinary focus » se distingue du focus méta-linguistique dont le domaine est la syllabe ou le phonème et qui pointe sur la forme phonologique du signifiant. Dans des études précédentes, nous avons montré, sur la base d'une analyse acoustique de F0, de l'intensité et de la durée, que les réalisations prosodiques du focus ordinaire (nouveauté ou contraste sur un élément lexical) et du focus méta-linguistique (focus syllabique) sont similaires. Les hauteurs de F0 et d'intensité sont identiques sur la première syllabe (pour un focus sur le mot ou directement sur la première syllabe), la pente qui ramène F0 de la première syllabe au niveau du F0 du mot, ont des décours différents [2]. Il s'agit donc de vérifier si cette apparente dissimilitude a une pertinence cognitive. Nous avons essayé de montrer par différentes expériences perceptives que trois types d'informations fonctionnelles sont véhiculées par la morphologie prosodique du focus[1] : (1) sa *Valence*, information grammaticale binaire sur la présence ou l'absence du focus sur l'élément, nous montrons qu'il est réalisé par un processus tonal (indice statique), (2) son *Intensité*, information pragmatique de quantité d'insistance, par exemple ou délivrer une valeur d'emphase ou encore pour instancier un choix entre deux éléments focalisés ; l'intensité est réalisé grâce une compétence psycho-acoustique de perception gradiente des variations de hauteurs tonales, montrée en particulier par Ladd, mais nous proposons que cette intensité fonctionnelle ne soit cognitivement pertinente qu'après que la valence positive de présence focus aie été instanciée, et non pas directement à partir d'un non focus comme le suggère l'expérience de Ladd et (3) son *Domaine*, information linguistique sur le segment qui constitue l'élément focalisé (la syllabe pour un focus méta-linguistique, le mot pour un focus de contraste ou de nouveauté), qui est réalisée par la valeur de la pente de F0 en fin de ton de focus (indice dynamique).

Nous rappelons ainsi brièvement les résultats d'une expérience de perception catégorielle, menée sur des stimuli « morphés » progessivement des contours sans focus aux contours avec focus, produits par un locuteur, et qui montre clairement une décision binaire par un effet magnet sur absence/présence de focus. Dans cette expérience, la valeur « frontière » de la présence de focus donne la base de la gradience qui opère au-dessus de la perception de présence de focus. Enfin, nous résumons deux expériences dont la première montre la compétence perceptive de discrimination du domaine (syllabe vs. mot) et dont la seconde, basée sur un morphing de la pente de F0, met en évidence un seuil de pente dans certains stimuli pour la perception du domaine de l'élément focalisé. Ces expériences nous permettent de proposer le modèle VID qui attribue à la morphologie prosodique du focus une fonction à trois valeurs informationnelles : Valence, Intensité, Domaine.

[1] « morphologie » de la prosodie est ici employé au sens générique, i.e. organisation de la forme prosodique

2. INDICES DE LA VALENCE DE LA FONCTION

Le corpus utilisé pour les expériences présentées ici (voir [2] et [1] pour plus de détails) est basé sur une même structure syntaxique porteuse, dans laquelle les items lexicaux varient uniquement selon la dimension phonotactique. Les phrases françaises enregistrées ont une longueur de 6 à 8 syllabes, et chaque item lexical comporte de 1 à 3 syllabes. Chaque phrase est enregistrée avec le focus réalisé alternativement sur chaque item lexical pour le focus sur le mot, sur chaque syllabe pour le focus métalinguistique et sans focus. Différents tests de perception [2] ont permis de valider la bonne perception du focus sur l'item visé. Le protocole de ces expériences est détaillé dans [1].

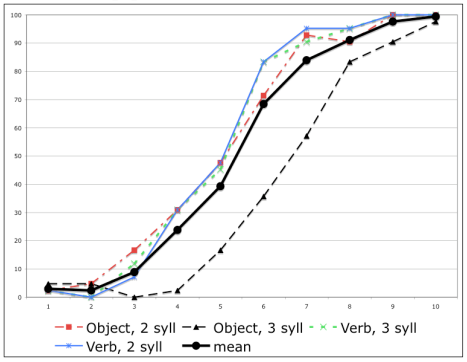


Figure 1. Pourcentages d'identification du focus pour les 4 continuums et pourcentage moyen. Les ordonnées représentent les 10 itérations du morphing depuis le stimulus neutre jusqu'au focalisé.

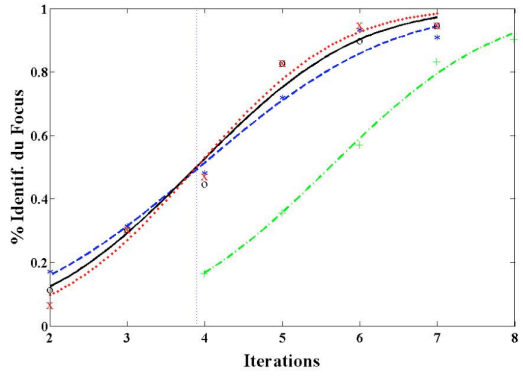


Figure 2. Interpolation Probit des résultats pour chaque continuum. Légende : (*) et tirets bleus : focus sur l'objet de 2 syllabes ; (+) et tirets-points verts : focus sur l'objet de 3 syl ; (o) et ligne continue : verbe de 3 syl ; (x) et pointillés : verbe 2 syl. La ligne verticale représente le seuil de perception du focus. En ordonnée l'itération et en indice le % d'identification du focus.

4 continuums acoustiques en 10 pas, partant de la phrase sans focus pour aboutir à la même phrase portant un focus sur un mot, ont été créés. Les pas de F0, d'intensité et de durée utilisés pour ces morphings sont tous inférieurs au seuil de perception, tel qu'il est décrit par Rossi [8]. Une tâche d'identification cherche à mettre en évidence un comportement quasi catégoriel pour la perception de ce continuum prosodique : un effet magnet [6]. Les résultats (cf. fig. 1) obtenus sont comparables à ceux décrits par Ladd [7]. Une régression logistique (analyse Probit) sur ces résultats (cf. fig 2) met en évidence le seuil de perception du focus aux alentours de la quatrième itération sur les 10.

Afin de compléter cette tâche d'identification, et de vérifier si le comportement est vraiment catégoriel, une expérience de discrimination a été réalisée, dont les résultats ne mettent en lumière aucun pic de discrimination aux alentours du seuil de perception. Nous devons donc conclure à un comportement de type effet magnet, avec une valeur de F0 (et aussi d'intensité) statique au-delà de laquelle les sujets perçoivent une focalisation (cf. fig. 3).

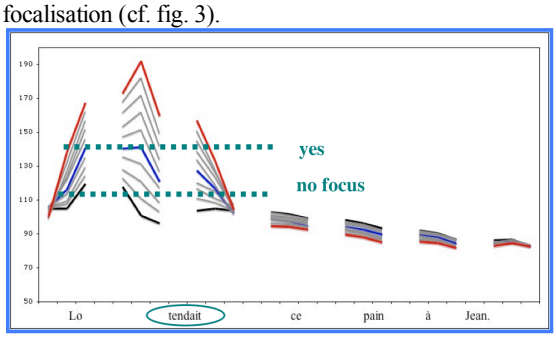


Figure 3. Morphologie de F0 dans le cas d'un verbe de 2 syllabes, pour les deux valences du focus. Les extrêmes correspondent aux deux références naturelles

3. INDICES DE L'INTENSITE DE LA FONCTION

Ladd [7] a montré qu'un auditeur en condition psycho-acoustique perçoit les variations de proéminence (F0 seulement) de manière gradiente, alors que les mêmes stimuli dans une tâche de perception linguistique sont perçus de manière catégorielle, conformément aux résultats présentés précédemment. Lors d'une expérience préliminaire, nous avons observé que les auditeurs utilisent ces capacités psychoacoustiques par exemple pour choisir un élément parmi deux éléments lexicaux focalisé au sein d'un même énoncé. Cette fonction, que appellerons fonction de préférence, supposent (1) une fonction déictique sur les deux éléments lexicaux, c'est-à-dire une valence positive de la fonction de focus sur les deux domaines lexicaux sur lesquels (2) va opérer l'intensité, par une gradience supérieure de l'un des deux pour implémenter la préférence. Ce genre de distinction est utilisé typiquement lors de dialogues homme-machine, pour exprimer la préférence du locuteur à propos d'un choix entre deux items. Les auditeurs écoutaient des phrases du type : « Préférez-vous une correspondance à *Paris* ou à *Londres* cette semaine ? », dans laquelle *Paris* ou *Londres* recevraient une gradience supérieure, les deux étant obligatoirement au-dessus du seuil de perception de la valence positive de focus. Pour être interprétable cette expérience demande à être reproduite avec des stimuli plus qui contrôlent systématiquement l'ordre de présentation, leur place dans l'énoncé (une même valeur variation de hauteur n'est pas perçue identiquement en début ou fin d'énoncé [3], et le domaine du focus, c'est-à-dire l'implémentation de la fonction de pointage lexicale vs. pointage méta-linguistique.

4. INDICES DU DOMAINE DE LA FONCTION

4.1. Expérience de discrimination

Lors de cette expérience, les auditeurs doivent juger si le locuteur parle plus spécifiquement d'une personne, d'une action ou d'un objet (en contraste avec un autre), ou s'il tâche de désambiguïser une syllabe mal comprise d'un mot particulier (focus métalinguistique). 25 sujets ont écouté tous les stimuli une fois seulement, et donné leur réponse grâce à l'interface présentée à la figure 4.

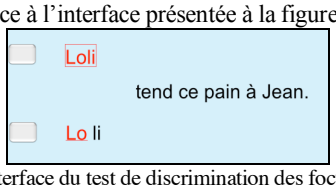


Figure 4. Interface du test de discrimination des focus sur le mot ou sur la première syllabe.

Les stimuli sélectionnés présentent toutes les longueurs et les positions possibles de focus sur un mot dans notre corpus, ainsi que leur pendant focalisé sur la première syllabe du même mot. Cela car en français, les mots focalisés sont réalisés avec une importante proéminence sur la première syllabe, et doivent donc être comparés aux stimuli ne portant un focus que sur cette syllabe-ci.

Table 1. Résultats de la tâche de discrimination.

	% bonnes réponses	
	brut	corrigé / hasard
Focus sur le mot	79,7	59,3
Focus 1ère syllabe	83,7	67
Total	81,7	63,3

Les auditeurs ont le sentiment de répondre au hasard, mais les résultats (cf. table 1) montrent un taux de discrimination largement supérieur au hasard, pour les deux types de focus.

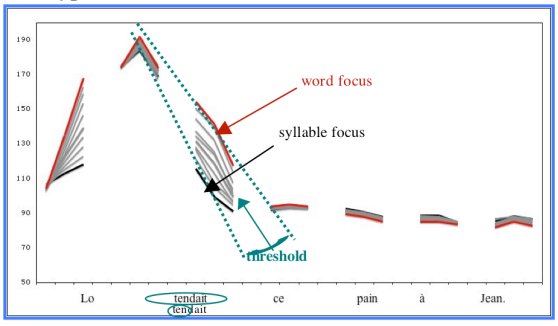


Figure 5. Morphologie de la F0 dans le cas des verbes de 2 syllabes, pour les deux domaines de focus. Les deux contours extrêmes correspondent aux deux références naturelles.

Il faut noter que l'analyse acoustique (cf. [2]) réalisée sur ces stimuli montre des niveaux de F0 et d'intensité similaire pour la première syllabe des mots réalisés avec un focus sur le mot ou sur la syllabe. Ceci est indirectement confirmé par l'expérience d'identification du focus présentée ci-après, pendant laquelle un seuil est relevé pour le focus sur le mot aussi bien que sur la syllabe. Cela pourrait indiquer que les indices permettant l'identification du domaine sont à rechercher dans la dynamique du contour du focus, entre la première syllabe et les suivantes (cf. figure 5). L'expérience suivante cherche à savoir si l'indentification du domaine correspond à un traitement catégoriel ou continu, et à déterminer la frontière morphologique dans le cas d'un traitement pseudo-catégoriel.

4.2. Expérience d'identification

Les stimuli utilisés pour cette expérience sont basés sur 6 phrases différentes, toutes construites sur la même structure syntaxique Sujet-Verbe-Objet. Un mot de chaque phrase porte le focus, soit le sujet, le verbe ou l'objet. La longueur des mots focalisés varie de 2 à 3 syllabes, tous les autres mots de la phrase étant monosyllabiques. Les phrases ont été enregistrées par un locuteur masculin francophone qui a produit le focus sur le mot en entier, ou bien seulement sur sa première syllabe. Un continuum est ensuite construit par analyse-synthèse, qui passe de l'une des phrases de chaque paire à la seconde grâce à un morphing des paramètres de F0, intensité et durée réalisé sur 10 pas (en utilisant le logiciel Praat pour modifier F0 et durée, et un script Matlab pour modifier l'intensité). Les différences introduites entre chacun de ces pas sont toutes en dessous du seuil de perception du glissando décrit par Rossi [8]. Cela donne finalement 6 continuums de 10 stimuli différents.

Table 2. ANOVA réalisée sur les réponses des sujets au test d'identification du focus sur le mot ou sur la syllabe. Les facteurs pris en compte sont le pas du morphing (Itération), la longueur du mot focalisé, la position de ce mot dans la phrase et les 3 répétitions.

Facteur	ddl	F	p	sig.
Itération	9	64,129	,000	*
Longueur	1	4,790	,053	
Position	2	9,981	,001	*
Répétition	2	1,067	,363	
Itération * Long	9	14,345	,000	*
Itération * Position	18	7,168	,000	*
Long * Position	2	1,785	,194	
Iter.*Long* Pos	18	3,599	,000	*

Les 11 sujets écoutent chaque stimulus, dans un ordre aléatoire. L'ensemble des stimuli est présenté trois fois à chaque sujet. Ils doivent répondre si le focus qu'ils perçoivent correspond à un focus sur le mot ou sur la première syllabe du mot.

La cohérence des résultats inter-sujets est validée par un alpha de Cronbach significatif (α=0.84). Ensuite, une ANOVA permet de mesurer l'influence relative des différents paramètres de ce test de perception sur les réponses des sujets : les différent pas du morphing, la longueur du mot focalisé, sa place dans la phrase (reliée à la fonction syntaxique du mot), et les trois répétitions. Les

effets principaux qui en ressortent (cf. table 2) permettent de souligner (1) l'effet primordial du morphing et (2) celui de la position du mot dans la phrase. Par ailleurs, (3) des interactions significatives sont relevées entre l'itération, la longueur du mot et sa position. Les trois répétitions n'ont aucun effet sur les réponses. Ces résultats confirment la capacité des auditeurs à percevoir le domaine du focus et donc à distinguer les deux fonctions sous-jacentes.

Cependant, l'effet important de la position du mot dans la phrase soulève d'autres questionnements sur l'influence de facteurs qu'il n'est pas possible de tester dans cette expérience, et nous conduit à vouloir poursuivre ces expériences plus avant. En particulier la nature phonologique des phonèmes (consonne voisée ou non) constituant la syllabe focalisée doit être contrôlée, et la position du mot focalisé découplée de sa fonction syntaxique. En particulier, on peut se demander si l'interaction significative entre la longueur du mot focalisé et l'itération est le reflet d'une plus grande facilité des sujets à traiter les mots quand ils sont plus longs. Mais comme la longueur n'est pas significative seule, et que les objets de deux syllabes reçoivent de bons résultats, cela pourrait aussi être dû à l'un des facteurs listés auparavant.

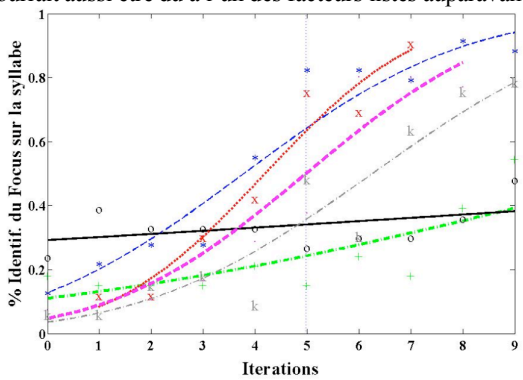


Figure 7. Résultats de l'analyse Probit. Les 6 courbes sont le résultat d'une interpolation des résultats moyens obtenus pour chaque continuum à chaque itération. Légende : (+) et tirés-points verts : sujet de 2 syllabes ; (*) et tirés bleus : objet de 2 syl ; (o) et ligne noire : verbe de 2 syl ; (k) et tirés-points gris : sujet de 3 syl ; (x) et pointillés rouge : objet de 3 syl ; (.) et tirés magenta : verbe de 3 syl. La ligne verticale indique la position de la frontière perceptive pour l'ensemble des courbes – sauf (b) et (c). En ordonnée l'itération et en indice le % d'identification du focus syllabique.

Afin de tester si les réponses des sujets sont de nature catégorielle ou non, une analyse Probit a été menée sur les résultats (cf. fig. 7). Cette analyse montre que 4 continuums sur les 6 testés montrent un passage abrupt des réponses de focalisation sur le mot à celles sur la syllabe, autour de la cinquième itération. Pour les deux autres, la focalisation syllabique n'a pas, ou mal, été reconnue. Ces résultats sont cohérents avec ceux de l'ANOVA, et nous poussent à poursuivre ces recherches afin de mieux comprendre les paramètres qui peuvent faire que certains stimuli sont bien ou mal reconnus. Le résultat important de cette étude reste cependant que pour 4 phrases, les sujets répondent à un continuum d'un focus sur le mot à un focus sur la syllabe comme s'il existait une frontière perceptive entre les deux.

5. CONCLUSIONS

Nous avons résumé ici plusieurs expériences perceptives qui confirment, après Ladd [7] que la prosodie du focus véhicule bien, à travers des indices statiques tonals traités dans une perception « quasi » catégorielle, la binarité présence ou absence de focus (Valence), mais ne peut cependant pas être réduite à cette seule valeur fonctionnelle, puisque des indices de gradience (traités seulement quand la catégorie également présence de focus est établie) informent sur l'Intensité du focus, qui ne peut pas exister en deça d'une valence positive (qui est utilisable dans des fonctions communicatives comme la préférence), tandis que des indices dynamiques informent sur le segment focalisé, à travers un effet « quasi » catégoriel, dont le seuil devra cependant être établi plus systématiquement. Le modèle VID, qui reprend les seuils établis dans les expériences perceptives, est en cours d'implémentation et de validation dans le système de synthèse, pour une mise en œuvre dans le système de dialogue de FT R&D.

REMERCIEMENTS

Nous remercions particulièrement Philippe Bretier et Franck Panaget de France Telecom R&D, pour les discussions fructueuses qui nous ont permis en particulier de cerner les fonctions communicatives du focus prosodique.

BIBLIOGRAPHIE

[1] Aubergé, V. 2001. Modélisation de la prosodie par formes globales : amont ou aval de la phonologie tonale ?. *23rd JEP* , France, 281-284.

[2] Brichet, C., Aubergé V., 2002. La prosodie de la focalisation en français : faits perceptifs. 94-99, 24es JEP, Nancy.

[3] Gussenhoven C (2003). Perceiving paralinguistic meaning. *Proceedings of Prosodic Interfaces 2003*, edited by Amina Mettouchi and Gäelle Ferré (eds).Université de Nantes. 47-49.

[4] Jackendoff, R. 2002. *Foundations of language*, Oxford: Oxford University Press.

[5] Kohler K. (2006). What is Emphasis and how is it coded ? Int Conf of Speech Prosody, Dresden.

[6] Kuhl, P. K. (1991). Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories; Monkeys do not. Perception & Psychophysics, 50, 93-107.

[7] Ladd., D.R. & Morton, R. 1997. The perception of intonational emphasis : continuous or categorical ? Journal of Phonetics, 25, 313-342.

[8] Rossi, M. 1978. La perception des glissandos descendants dans les contours prosodiques. *Phonetica*, 35, 11-40.

[9] Rossi, M. 1985. L'intonation et l'organisation de l'énoncé, *Phonetica,* 42, 135-153.

Session VII

Poster

Mardi 13 juin 2006 - 11h45 12h30

Représentation acoustique compacte pour un système de reconnaissance de la parole embarquée

Christophe Lévy, Georges Linarès, Jean-François Bonastre

Laboratoire Informatique d'Avignon
339 chemin des meinajaries, BP 1228, 84911 Avignon, France
{christophe.levy, georges.linares, jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

Speech recognition applications are known to require a significant amount of resources (training data, memory, computing power). However, the targeted context of this work -mobile phone embedded speech recognition system- only authorizes few resources. In order to fit the resource constraints, an approach based on a HMM system using a GMM-based state-independent acoustic modeling is proposed in this paper. A transformation is computed and applied to the global GMM in order to obtain each of the HMM state-dependent probability density functions.

The proposed approach is evaluated on a French digit recognition task. Our method leads a Digit Error Rate (DER) of 2%, when the system respects the resource constraints. Compared to an HMM with comparable resource, our approach achieved a DER relative decrease more than 50%.

1. INTRODUCTION

Le marché de la téléphonie mobile n'a cessé de croître depuis le milieu des années 90. D'après l'ARCEP[1] au troisième trimestre 2005 ce marché atteignait les 46 millions d'abonnements (plus de 73% de la population française - enfants compris - possèderait un téléphone).

Les téléphones de dernière génération offrent de nouvelles fonctionnalités : gestion des rendez-vous, outils de bureautique, jeux,*etc.* Certains modèles résultant de la fusion entre un téléphone et un organiseur numérique (PDA) offrent même l'ensemble des fonctionnalités de ces deux appareils. Ajouté à cela, une course à la miniaturisation a amené à des outils (téléphone, PDA, PDAphone, *etc.*) de taille très réduite. L'intégration de la Reconnaissance Automatique de la Parole (RAP) dans ces systèmes devient un enjeu majeur en terme d'ergonomie.

Les performances obtenues par les systèmes de RAP actuels permettent d'envisager des applications réelles. Ces systèmes sont généralement basés sur une modélisation stochastique d'unités acoustiques (mots, phonèmes, diphones, tri-phones, *etc.*). Cette approche probabiliste nécessite beaucoup de données pour l'apprentissage des modèles acoustiques, un espace de stockage conséquent pour les modèles et une puissance de calcul non négligeable pour la reconnaissance. Ces points deviennent cruciaux dès lors que le moteur de RAP est intégré dans un système embarqué, ces derniers ne possèdent que très peu de ressources (mémoire et/ou puissance de calcul).

[1] Autorité de Régulation des Communications Electroniques et des Postes - `http://www.art-telecom.fr/`

Dans cet article, nous essayons d'apporter une réponse à ces problèmes. Pour cela, nous proposons une appoche basée sur un HMM (Hidden Markov Model). Notre proposition consiste à représenter l'espace acoustique dans un seul modèle générique (un GMM, Gaussian Mixture Model) puis à dériver les modèles des différents phonèmes depuis ce modèle générique, en mutualisant une large partie des modèles (*cf.* figure 1). Une première ébauche de cette approche a été présentée dans [6], une version détaillée peut être trouvée dans [7].

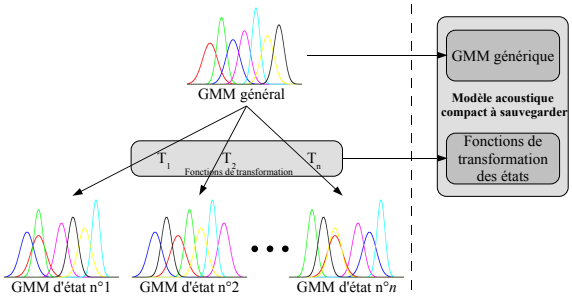


FIG. 1: principe général de l'approche proposée.

Dans la littérature différentes méthodes pour intégrer un moteur de RAP dans des systèmes embarqués sont proposées [1, 10, 9]. Les deux derniers travaux cités précédemment réduisent *a posteriori* la taille des modèles en mutualisant les paramètres par une approche ascendante, qui part d'un ensemble de modèles spécifiques à chaque unité acoustique et les regroupe hiérarchiquement en partageant les paramètres les plus proches. Notre approche vise les mêmes objectifs, mais propose une démarche descendante qui modélise l'environnement acoustique par un unique modèle générique, puis dérive celui-ci pour obtenir les modèles correspondant à chaque unité acoustique. L'intérêt d'une approche descendante réside dans une meilleure représentation de l'espace acoustique global étant donné que le modèle générique est appris sur l'ensemble des données d'apprentissage.

Après une brève présentation des deux corpus utilisés (section 2), nous détaillons l'approche proposée dans la section 3. Les principaux résultats sont présentés dans la section 4. Pour finir, quelques conclusions et perspectives sont présentées dans la section 5.

2. CORPUS

Afin d'évaluer l'approche proposée dans cet article, deux corpus ont été utilisés. Le premier, BREF120 [5], a servi uniquement à l'apprentissage du modèle générique. Le se-

cond corpus, BDSONS [3], permet de construire les modèles de chaque unité acoustique (ici les phonèmes). Il est également utilisé pour les expériences de reconnaissance.

2.1. BREF120

Le corpus BREF120 est composé de parole lue issue du journal français *Le Monde*. Il est constitué d'une centaine d'heures de parole enregistrées par 120 personnes (55 hommes et 65 femmes).

Ce corpus a servi à l'apprentissage du modèle acoustique générique. Ce modèle a ensuite été adapté au corpus de test, en utilisant le sous-ensemble ADAPT_SET du corpus BDSONS (*cf.* la section 2.2).

2.2. BDSONS

Le corpus BDSONS est composé de phrases phonétiquement équilibrées, de suites CVC, de logatomes, de chiffres, *etc.* Ces séquences ont été prononcées par 32 locuteurs (16 hommes et 16 femmes).

Seul le sous-corpus contenant les chiffres isolés a été retenu (composé de 15 hommes et 15 femmes). Ce sous-corpus a été divisé en deux sous-ensembles :
– un pour l'adapation (ADAPT_SET). Il contient 700 occurences de chiffre prononcées par 7 locuteurs (4 hommes et 3 femmes). Ce sous-ensemble a été utilisé pour adapter le modèle acoustique générique. Il a également été utilisé pour construire les modèles correspondant aux phonèmes.
– un pour l'évaluation (TEST_SET). Il est composé de 2300 occurences de chiffre prononcées par 23 locuteurs (11 hommes et 12 femmes).

Les locuteurs du corpus BDSONS sont différents de ceux du corpus BREF. Les locuteurs des corpus ADAPT_SET et TEST_SET sont aussi différents. Les tests sont donc réalisés en mode indépendant du locuteur, les locuteurs de test n'apparaissant jamais durant les phases d'apprentissages.

L'objectif de notre travail est de proposer un système de RAP embarqué généraliste, capable de reconnaitre des commandes vocales, des noms propres *etc.* Nous utilisons donc une modélisation en unités acoustiques plutôt qu'une représentation par mots.

Cependant, le corpus de test étant limité à des chiffres isolés, les résultats sont exprimés en DER (Digit Error Rate - taux d'erreur de reconnaissance de chiffre).

3. APPROCHE BASÉE SUR UN MODÈLE GÉNÉRIQUE

L'approche proposée dans cet article consiste à apprendre un modèle générique avec un ensemble suffisant de données puis à dériver ce modèle pour chaque unité acoustique, à l'aide d'une transformation basique dont seuls les paramètres seront stockés pour caractériser le modèle d'un phonème donné. Les modèles utilisés (modèle générique et modèles des états des HMMs) sont des GMM. La transformation utilisée consiste en une ré-estimation des poids suivie d'une sélection des *N-Best* gaussiennes (seulement les *N-Best* gaussiennes sont sauvegardées). Cette transformation est appelée WRE (Weight Re-Estimation) dans cet article. Elle peut être précédée d'une adaptation linéaire globale du GMM (avant l'adaptation des poids). Durant cette phase, la transformation est la même pour toutes les gaussiennes. Cette phase d'adaptation est nommée ULT (Unique Linear Transformation) dans la suite de cet article.

3.1. Ré-estimation des poids : WRE

Cette étape consiste à adapter les poids du GMM générique afin d'obtenir le GMM dépendant de l'état. Les états sont donc différenciés entre eux uniquement par le vecteur de poids du GMM.

Pour la ré-estimation des poids, deux méthodes sont proposées :
– la première basée sur le critère de maximum de vraisemblance (MLE),
– la seconde basée sur une approche discriminante (MMIE).

Cette étape ne nécessite que très peu de ressource. En effet, pour stocker le modèle acoustique il suffit de stocker le GMM générique et un vecteur de poids. De plus, Tous les poids n'ont pas été conservés, seuls ceux correspondant aux *N-Best* gaussiennes ont été conservés, de manière à réduire l'espace nécessaire au stockage des modèles acoustiques. Pour le calcul de la vraisemblance, il suffit de calculer la vraisemblance de chaque gaussienne du GMM générique puis d'effectuer une somme pondérée en fonction du vecteur de poids considéré.

• **MLE** :

Le poids de la $i^{ème}$ gaussienne (w_i) est ré-estimé suivant le critère de maximum de vraisemblance (le critère MAP - *Maximum A Posteriori* [4] - pourrait être utilisé mais devant la faible quantité de données disponibles, nous avons préféré utiliser un critère MLE). La fonction de ré-estimation utilisée est donc :

$$w'_i = \frac{w_i * Vrais(tr|g_i)}{\sum_{g_j=1}^{nb_g} w_j * Vrais(tr|g_j)} o$$

Vrais(tr|g$_x$) correspond à la vraisemblance de données relatives à l'état (tr) pour la gaussienne g_x.

• **MMIE** :

L'apprentissage des HMM en utilisant un critère discriminant maximisant l'information mutuelle (Maximum Mutual Information Estimation - MMIE) a déjà été largement étudié, notamment dans [2].

Dans [7], une technique rapide pour ré-estimer les poids suivant un critère MMIE a été présentée. La règle de mise à jours des poids proposée est :

$$w'_{jm} = w_{jm} * \frac{w_{jm}}{\sum_k w_{km}}$$

où w_{jm} représente le poids de la $j^{ème}$ gaussienne de l'état m.

3.2. Transformation linéaire : ULT

La méthode LIAMAP présentée dans [8] permet d'adapter globalement un GMM en utilisant uniquement une transformation simple. Cette transformation nous permet

d'adapter le GMM générique (*GMMgnl*) avec les données propres à chaque état. La transformation porte sur les paramètres de moyenne et de variance. La forme générale de cette transformation est donnée ci-dessous :

$$\mu_{GMMetat} = \alpha * \mu_{GMMgnl} + \beta$$

$$\Sigma_{GMMetat} = \alpha^2 * \Sigma_{GMMgnl}$$

avec α (commun pour $\mu_{GMMetat}$ et $\Sigma_{GMMetat}$) et β définis ci-après.

L'idée principale de cette adaptation (*cf.* figure 2) est d'estimer une transformation entre les gaussiennes (μ, Σ) et $(\tilde{\mu}, \tilde{\Sigma})$ obtenues :

1. en fusionnant toutes les gaussiennes du GMM générique afin d'obtenir (μ, Σ).
2. en adaptant (avec MAP) le GMM générique avec les données spécifiques de l'état et en fusionnant ensuite toutes les gaussiennes de ce nouveau GMM afin d'obtenir $\widetilde{\mu}$ and $\widetilde{\Sigma}$.

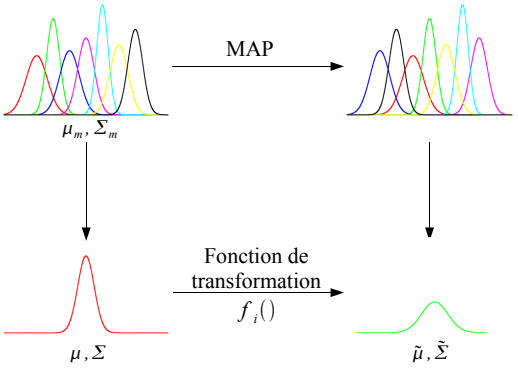


FIG. 2: LIAMAP : principe général de la transformation LIAMAP.

Chaque gaussienne finale (définie par sa moyenne μ'_m et sa matrice de covariance Σ'_m) est calculée de la manière suivante :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}(\mu_m - \mu) + \widetilde{\mu} \quad (1)$$

$$\Sigma'_m = \widetilde{\Sigma}\Sigma^{-1}\Sigma_m \quad (2)$$

L'équation 2 peut-être développée en :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu_m - \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \quad (3)$$

Posons

$$\alpha = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2} \quad (4)$$

et

$$\beta = -\widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \quad (5)$$

Les équations 2 et 3 deviennent :

$$\mu'_m = \alpha\mu_m + \beta \quad (6)$$

et

$$\Sigma'_m = \alpha^2\Sigma_m \quad (7)$$

Les équations 7 et 8 montrent que nous obtenons une simple transformation linéaire définie par les vecteurs α et β (la transformation est partagée par l'ensemble des gaussiennes du GMM).

ULT est ici présentée comme une première étape (optionelle) avant la ré-estimation des poids. L'étape suivante

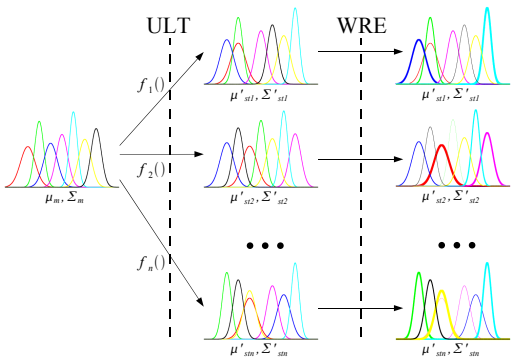


FIG. 3: Transformation permettant de passer du GMM général, indépendant des états, au GMM dépendant des états.

(WRE, *cf.* 3.1) est toujours appliquée. La figure 3 illustre le processus complet.

Durant le test, cette étape nécessite le calcul du GMM dépendant de l'état avant de pouvoir calculer la vraisemblance d'un état. Elle impose aussi de calculer la vraisemblance de chaque gaussienne de chaque état contrairement à une approche WRE simple (sans ULT préalable).

3.3. Evaluation de l'occupation mémoire

Nous avons fixé deux limites d'occupation mémoire proches des capacités réelles disponibles dans un téléphone portable.

Pour chaque approche - WRE et ULT+WRE - nous avons estimé la taille des modèles acoustiques en terme de nombre de paramètres. Ceci permet de fixer le nombre de gaussiennes pour chacune des trois approches afin que chaque approche nécessite la même quantité de mémoire.

4. RÉSULTATS

Les expériences ont été effectuées sur une tâche de reconnaissance de chiffre isolé avec le corpus BDSONS. A titre de comparaison, nous donnons aussi les résultats par un système HMM classique[2] respectant les contraintes mémoire que nous nous sommes fixées.

Le tableau 1 résume l'ensemble des expériences effectuées. Il présente les performances en terme de taux d'erreur de reconnaissance (Digit Error Rate - DER) de chaque méthode.

L'approche présentée dans cet article (ULT+WRE ou WRE uniquement) permet une diminution importante du DER. La réduction, relative, du DER varie entre 9% (WRE/MMIE limite mémoire inférieure) et 53% (WRE/MMIE limite mémoire supérieure). Le DER atteint 2,09% avec notre meilleure approche alors qu'il se situe aux alentours de 4,43% pour l'approche HMM classique[3].

Le gain, en terme de DER, apporté par l'étape ULT est à nuancer au regard du surcoût de calcul qu'impose cette

[2]Le HMM est appris avec le corpus BREF puis une adaptation MAP est effectuée avec les données ADAPT_SET de BDSON.

[3]un modèle HMM classique sans contrainte mémoire avec 39 coefficient PLP et 128 gaussiennes par état (non-contextuel) obtient un DER de 0,96%.

TAB. 1: DER pour les approches WRE et ULT+WRE (2700 tests).

Model size	HMM	WRE/MLE	WRE/MMIE	ULT+WRE/MLE	ULT+WRE/MMIE
6k	4,96%	4,17%	4,52%	3,39%	3,22%
11k	4,43%	3,09%	2,09%	3,00%	2,70%

étape. En effet, l'approche WRE est très économe en puissance de calcul, elle permet de calculer l'ensemble des vraisemblances pour une observation donnée (une vraisemblance par état des machines phonétiques) en ne calculant qu'une seule fois la vraisemblance des gaussiennes du modèle générique. Les vraisemblances pour chaque état correspondent simplement à une somme pondérée des vraisemblances calculées avec le modèle générique. A contrario, l'approche ULT+WRE nécessite de calculer le GMM de l'état puis les vraisemblances de chacune des composantes avant d'obtenir le résultat par une somme pondrée de ces dernières vraisemblances. Il convient donc de choisir entre ULT+WRE et WRE en fonction des ressources (d'un point de vue puissance de calcul) disponibles.

5. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté une approche permettant d'intégrer un système de reconnaissance automatique de la parole dans un système embarqué (téléphone portable, PDA, *etc.*).

La technique présentée dans ce papier est basée sur le formalisme des HMM (classiquement utilisé en RAP). Nous proposons une optimisation des HMM en terme de ressources mémoire et de calcul qui permet d'atteindre un niveau de performance intéressant dans le contexte applicatif des systèmes embarqués.

Les résultas présentés dans cet article montrent que notre approche permet une réduction relative des taux d'erreur variant entre 9% (WRE/MMIE, petit modèle) et 53% (WRE/MLE, grand modèle) comparée à un modèle HMM classique. Par exemple, notre méthode permet un taux d'erreur à 2,09% pour les plus gros modèles, à comparer au résultat du HMM classique, 4,43% (pour comparaison, le système HMM classique obtient un taux d'erreur avoisinant les 1% avec un modèle acoustique composé de plus d'un million de paramètres).

Dans l'étude présentée aucune adaptation, au locuteur ou à l'environnement n'a été réalisée. Une suite logique de ce travail serait d'envisager différentes formes d'adaptation. Une approche prometteuse consiste à réaliser cette adaptation au niveau du GMM générique. En effet, en se basant sur les méthodes proposées en reconnaissance du locuteur, quelques trames peuvent suffir à adapter ce modèle au locuteur (ou à l'environnement), indépendamment des mots prononcés. Dans le cadre de l'adaptation ULT, un second niveau d'adaptation pourrait être envisagé, agissant au niveau de chaque état.

RÉFÉRENCES

[1] S. Astrov, J.G. Bauer, and S. Stan. High performance speaker and vocabulary independent ASR technology for mobile phones. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2003)*, pages 281–284, Hong Kong, April 2003.

[2] L.R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model parameters for speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, pages 49–52, Tokyo, Japan, April 1986.

[3] R. Carré, R. Descout, M. Eskénazi, J. Mariani, and M. Rossi. The French language database : defining, planning and recording a large database. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1984)*, pages 324–327, San Diego, California, USA, March 1984.

[4] J.L. Gauvain and C.H. Lee. Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains. In *IEEE Transactions on Speech and Audio Processing*, volume 2-2, pages 291–298, April 1994.

[5] L.F. Lamel, J.L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'1991)*, pages 505–508, Gênes, Italie, September 1991.

[6] C. Lévy, G. Linarès, and J.F. Bonastre. Mobile phone embedded digit-recognition. In *Workshop on DSP in Mobile and Vehicular Systems*, Sesimbra, Portugal, September 2005.

[7] C. Lévy, G. Linarès, P. Nocera, and J.F. Bonastre. *Embedded mobile phone digit-recognition*, chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2. Springer Science, H. Abut, J.H.L. Hansen and K. Takeda edition, 2006.

[8] D. Matrouf, O. Bellot, P. Nocera, Linarès, and J. F. Bonastre. Structural linear model-space transformations for speaker adaptation. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'2003)*, pages 1625–1628, Geneva, Switzerland, September 2003.

[9] J. Park and H. Ko. Compact acoustic model for embedded implementation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'2004)*, pages 693–696, Jeju Island, Korea, October 2004.

[10] S.J. Young. The general use of tying in phoneme-based HMM speech recognisers. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1992)*, pages 569–572, San Francisco, California, USA, March 1992.

Mesures de confiance trame-synchrone

Joseph Razik, Odile Mella, Dominique Fohr et Jean-Paul Haton

Loria - UMR 7503 - Equipe Parole
Vandœuvre-Lès-Nancy, France
prenom.nom@loria.fr http ://parole.loria.fr

ABSTRACT

This paper presents some confidence measures for large vocabulary speech recognition that can be evaluated directly within the first steps of the recognition process. Having some clues to drive the recognition process may help to improve the accuracy of a sentence. Confidence measures may fit to this goal, so we propose some measures that can help the recognition as early as possible, without having to wait for the recognition process to be completed. Furthermore, our confidence measures are local, and they are based on partial word graphs. Experiments on a French broadcast news corpus are presented, and give results close to the post calculated version of the measures.

1. INTRODUCTION

Dans la plupart des systèmes de reconnaissance automatique de la parole, l'utilisation d'une mesure de confiance associée à un mot reconnu se fait lors d'une étape postérieure distincte. Ceci est principalement dû à leur objectif : validation dans un processus de recherche de mots clés [2], sélection de mots corrects en apprentissage non supervisé [8] ou bien encore détection de mots hors vocabulaire [3] .

Or, il serait parfois intéressant d'avoir accès à une mesure de confiance directement pendant le processus de reconnaissance et donc au sein même du moteur de reconnaissance. Ainsi, le moteur pourrait tenir compte d'un effet de *doute* sur des mots et modifier le processus de reconnaissance en conséquence.

Nous nous plaçons dans ce cadre en proposant dans cet article des mesures intégrables directement dans le moteur de reconnaissance et calculables quasiment de manière synchrone avec la progression de la reconnaissance. En effet, la plupart des mesures nécessitent le décodage complet de la phrase à reconnaître pour prendre une décision ou donner une estimation. Nous utilisons comme moteur de reconnaissance Julius [5], un moteur grand vocabulaire effectuant un traitement en deux passes. La première passe détermine pour chaque trame un nombre restreint de mots hypothèses. La deuxième passe fournit la phrase finale ayant la plus grande vraisemblance. Les mesures que nous proposons se placent au niveau du mot ; elles ont un caractère local et interviennent directement au cours du déroulement de la première passe.

Plusieurs approches ont été proposées afin de calculer des mesures de confiance, comme la mise en compétition de modèles, une comparaison avec le résultat d'un décodeur phonétique [4] ou l'utilisation de paramètres heuristiques (nombre de phonèmes dans le mot, durée des phonèmes, etc.). D'autres méthodes tentent d'estimer la probabilité *a posteriori* des mots reconnus ou bien sont basées sur les informations issues d'un graphe de mots contenant les multiples chemins qui vont du début à la fin de la phrase (Ortmanns [6]). Les mesures présentées dans cet article font partie de cette dernière approche. Plus précisément, elles sont inspirées de mesures à caractère local proposées dans nos travaux précédents [7], et d'une mesure fondée sur la probabilité *a posteriori* et décrite par F. Wessel et al. [10].

La section 2 décrit les mesures de confiance développées (mesures locales et mesure de la probabilité *a posteriori*). La section 3 décrit les conditions d'expérimentation ; elle contient une description rapide du moteur de reconnaissance, de sa structure interne (dans laquelle nous puisons les informations pour les différentes mesures) et des corpus qui sont utilisés. Pour finir, les différents résultats et tests sont décrits en section 4.

2. MESURES DE CONFIANCE

Le but d'une mesure de confiance est en général de donner une estimation de la probabilité qu'un mot reconnu soit correct. Les mesures décrites dans cet article permettent de donner cette estimation directement au cours du processus de reconnaissance du moteur, pendant le déroulement de la première passe de celui-ci. Pendant cette passe, le moteur génère une structure interne contenant pour chaque trame un nombre restreint d'hypothèses de mots possibles. Cette structure servira au cours de la deuxième passe à déterminer la phrase la plus vraisemblable. Afin de pouvoir fournir une mesure de confiance pour les mots pendant le déroulement de la première passe, les mesures présentées ont un caractère local : elles n'utilisent que des informations disponibles au moment du calcul. Nous présentons dans cette section des mesures de confiances locales. Toutes sont basées sur le graphe de mots interne au moteur de reconnaissance, tandis qu'une seule est issue de la mesure décrite par F. Wessel et al. [10] estimant la probabilité *a posteriori*.

Introduisons quelques notations : soit w un mot hypothèse, τ son instant de début et t son instant de fin. Une phrase commence au temps 1, se termine au temps T et x_1^T représente la séquence d'observations du temps 1 au temps T. Soit $[w, \tau, t]$ un mot hypothèse spécifique, et $[w, \tau, t]_1^M$ une séquence de M mots $[w_i, \tau_i, t_i]$, où $\tau_1 = 1, t_M = T$ et $t_{i-1} = \tau_i - 1$ pour $i = 2, \ldots, M$. $C([w, \tau, t])$ représente la mesure de confiance pour le mot hypothèse $[w, \tau, t]$.

2.1. Mesures de confiance locales

Pour concevoir nos mesures de confiance, nous utilisons un graphe de mots extrait du treillis d'exploration du moteur de reconnaissance. L'idée est de pouvoir calculer pendant la phase de reconnaissance du moteur une mesure de confiance pour chaque mot de la phrase.

Mesure basée sur des probabilités unigrammes Cette mesure utilise seulement des informations très simples et très locales : les scores acoustiques et les probabilités unigrammes. Cette mesure est similaire à un rapport de vraisemblance [9] entre le mot analysé et d'autres mots hypothèses, mais nous ne prenons en compte dans le rapport que les mots ayant survécu à l'élagage du faisceau de recherche et satisfaisant des conditions sur leur temps de début, de fin, et sur leur durée. Pour cette mesure, nous utilisons le graphe de mots interne du moteur de reconnaissance afin de sélectionner les mots hypothèses. Pour cela, nous introduisons un facteur de relâchement sur les contraintes temporelles de sélection de mots dans le graphe. Ces contraintes concernent les temps de début et de fin, et la longueur des mots hypothèses. Par exemple, pour un mot hypothèse $[w, \tau, t]$ et un taux de relâchement de 0,5, nous considérons les mots qui apparaissent avec un temps de début égal à $\tau \pm 50\%$ de la longueur de w. Le temps de fin et la longueur du mot sont traités de la même manière. Or, avec ce relâchement de contraintes, plusieurs occurrences du même mot hypothèse satisfaisant ces nouvelles contraintes peuvent apparaître. Dans ce cas, l'hypothèse ayant obtenu le score acoustique maximal est retenue. Nous introduisons également des facteurs d'échelle, à la fois pour le score acoustique (α) et pour le score du modèle de langage (β). Notre première mesure est ainsi définie par l'équation Eq. 1.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha . p(w)^\beta}{\sum_{[w',\tau',t']\in E} max(p(x_{\tau'}^{t'}|w'))^\alpha . p(w')^\beta} \quad (1)$$

où E est l'ensemble des mots qui satisfont les contraintes de temps et de longueur données par le facteur de relâchement.

Mesures basées sur des probabilités bigrammes La mesure précédente est modifiée afin de prendre en compte des informations sur le voisinage du mot par l'intermédiaire de probabilités bigrammes. Ces probabilités pour un mot $[w, \tau, t]$ sont calculées avec tous les mots précédents w_p qui se terminent au temps $\tau - 1$. Nous obtenons ainsi l'équation Eq. 2.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha \sum_{w_p} \big(p(w|w_p)p(w_p)\big)^\beta}{\sum_{[w',\tau',t']\in E} max(p(x_{\tau'}^{t'}|w'))^\alpha \sum_{w'_p} \big(p(w'|w'_p)p(w'_p)\big)^\beta} \quad (2)$$

Les informations apportées par la probabilité bigramme peuvent ne pas être suffisantes car encore trop locales. C'est pourquoi nous collectons encore un peu plus d'informations en utilisant les probabilités bigrammes avec les mots *précédents* et *suivants*. Pour un mot hypothèse $[w, \tau, t]$, nous considérons tous les mots précédents possibles w_p qui se terminent à $\tau - 1$, et tous les mots suivants possibles w_s qui commencent à $t + 1$. Nous introduisons une notation supplémentaire Γ, qui représente pour un mot hypothèse $[w, \tau, t]$ les informations issues des probabilités bigrammes :

$$\Gamma_{[w,\tau,t]} = \sum_{w_p} \sum_{w_s} \{p(w|w_p).p(w_s|w).p(w_p)\}^\beta \quad (3)$$

Nous définissons notre mesure par l'équation Eq. 4.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha \Gamma_{[w,\tau,t]}}{\sum_{[w',\tau',t']\in E} max(p(x_{\tau'}^{t'}|w'))^\alpha \Gamma_{[w',\tau',t']}} \quad (4)$$

2.2. Les mesures basées sur la probabilité a posteriori

La probabilité *a posteriori* est un bon indicateur de l'exactitude d'un mot et beaucoup de mesures de confiance s'appuient sur cette probabilité. F. Wessel et al. [10] ont proposé une mesure de confiance définie par la probabilité *a posteriori* d'un mot. Leur méthode pour calculer cette probabilité est inspirée de l'algorithme *forward-backward*, mais appliqué cette fois avec la granularité du mot. Cet algorithme est appliqué à un graphe de mots semblable à celui généré par le moteur de reconnaissance. La mesure définie par F. Wessel et al. [10] nécessite que le graphe de mots soit totalement généré pour pouvoir déterminer la probabilité *a posteriori*. Ainsi, une utilisation directe pendant le processus de reconnaissance n'est pas possible.

Nous proposons alors de modifier cette mesure afin d'obtenir une mesure plus locale. L'idée est simple : considérer pour chaque mot, non pas le graphe entier, mais un sous graphe contenant le mot à analyser. Pour chaque mot, nous déterminons un voisinage centré sur celui-ci, délimitant une plage temporelle à partir de laquelle nous extrayons un sous-graphe du graphe total. Puis, ce sous graphe est vu comme le graphe de mots associé à une *pseudo phrase* équivalente à une sous-séquence de la phrase. Afin de déterminer ce voisinage, le calcul de la mesure est en retard par rapport à la progression temporelle du moteur de reconnaissance, mais seulement de quelques trames.

Pour ces mesures basées sur la probabilité *a posteriori*, nous avons également utilisé des facteurs d'échelle et un facteur de relâchement. Il est à noter que les mesures sont calculées avec des probabilités bigrammes.

3. CONDITIONS D'EXPÉRIMENTATION

Pour chacune des différentes mesures précédemment définies, les conditions d'expérimentation sont identiques, et ces mesures ont accès exactement aux mêmes données. Nous décrivons dans cette section les différents paramètres définissant ces conditions d'expérimentation.

3.1. Les modèles acoustiques, de langage et le lexique

Le corpus d'apprentissage des modèles acoustiques utilisés par le système de reconnaissance se compose de 7 heures de bulletins d'informations radiophoniques, contenant uniquement de la parole large bande (pas de téléphone, pas de musique pure et pas de parole sur fond musical). Le signal est paramétré par des MFCC en appliquant une normalisation MCR (Mean Cepstral Remo-

val). Chaque phonème est modélisé à l'aide d'un modèle HMM.

Le modèle de langage a été appris par l'intermédiaire du CMU Toolkit [1] sur 16 ans du journal français « Le Monde », complété par une transcription manuelle de tout le corpus d'apprentissage de bulletins d'informations radiophoniques. Finalement, nous avons 2.5M de bigrammes et 5.8 M de trigrammes.

Le lexique de 54747 mots contient à la fois des mots au sens habituel du terme, mais aussi des groupes de mots. En effet certains mots ont été regroupés en une seule entité dans le lexique et ne comptent donc que pour un *mot*. Par exemple, la séquence « de la » est représentée par une seule entité « de_la », mais aussi certains noms comme « Aix_les_Bains ».

3.2. Le moteur de reconnaissance Julius

Julius [5] est un système de reconnaissance de la parole grand vocabulaire. Le processus de reconnaissance s'effectue en deux passes : une première passe trame-synchrone qui génère un treillis d'exploration en utilisant un modèle de langage bigramme, et une deuxième passe utilisant ce treillis et des modèles trigrammes pour aboutir à la phrase reconnue. Nous nous servons de Julius dans sa version 3.4.1-multipath, compilée avec l'option v2.1 pour une précision accrue.

Le treillis d'exploration Cette structure interne du moteur de reconnaissance, générée pendant la première passe, donne accès pour chaque trame du signal à plusieurs informations : les hypothèses de mots pouvant se terminer à cette trame, le mot précédent, leur score acoustique et de modèle de langage, etc. Nous obtenons en moyenne 470 mots hypothèses par trame avec un maximum de 2523 mots. En fait, le treillis d'exploration peut être considéré comme un graphe de mots.

3.3. Le corpus de développement et de test

Un corpus, également constitué de bulletins d'informations radiophoniques mais indépendant de celui utilisé pour l'apprentissage des modèles acoustiques, a été divisé en deux parties : une pour le développement, et une pour les tests. Le corpus de développement, d'une durée de 56 minutes, sert à mettre au point le seuil de décision et les facteurs d'échelle des mesures. Le corpus de test est d'une durée de 53 minutes. Ces corpus contiennent respectivement 12135 et 11272 mots. Le taux de reconnaissance moyen sur les deux corpus est d'environ 70,9%. L'ensemble du corpus est constitué de parole large bande, sans parole téléphonique ni musique, mais des phrases peuvent contenir un bruit ou une musique de fond. Le nombre moyen de *mots* par phrase du corpus de test est de 11,5.

3.4. L'évaluation

Pour évaluer les différentes mesures, nous étiquetons les mots de la phrase en deux classes : acceptation et rejet. Cet étiquetage dépend d'un seuil qui définit une frontière entre ces deux classes. Les étiquettes des mots sont ensuite comparées aux fichiers de référence. Ainsi, nous pouvons évaluer deux taux : le taux de *Fausse Acceptation* (FA) et le taux de *Faux Rejet* (FR). Le taux de fausse acceptation correspond aux cas où un mot incorrect est accepté, et le taux de faux rejet correspond aux cas où un bon mot est rejeté.

$$FA = \frac{nb.\,de\,mots\,incorrects\,étiquetés\,Acceptation}{nb.\,de\,mots\,incorrects} \quad (5)$$

$$FR = \frac{nb.\,de\,mots\,corrects\,étiquetés\,Rejet}{nb.\,de\,mots\,corrects} \quad (6)$$

A l'aide de ces deux taux et avec différents seuils de confiance, nous pouvons représenter la courbe DET (Detection-Error Tradeoff) et en déduire le taux EER d'égale erreur (Equal Error Rate).

4. TESTS ET RÉSULTATS

Pour nos tests, nous utilisons plusieurs valeurs pour le facteur de relâchement : 0,1 ; 0,2 ; 0,3 et 0,5. Concernant les facteurs d'échelle, nous considérons deux couples de valeur : $(\alpha;\beta) = (1;1)$ et $(\alpha;\beta) = (0,1;0,95)$. Le deuxième couple (0,1 ; 0,95) correspond aux valeurs optimales pour la mesure de confiance de F. Wessel et al. [10] obtenues sur notre corpus de développement. Des expérimentations supplémentaires montrent que ce couple de valeur est également optimal pour les autres mesures définies. Dans tous les tableaux qui suivent, les valeurs représentent le taux EER.

Dans le premier test, nous considérons notre mesure simple (Eq. 1) qui ne repose que sur les scores acoustiques et les probabilités unigrammes.

TAB. 1: Taux d'EER avec la probabilité unigramme (Eq. 1).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	39,9%	41,5%	43,5%	45,5%
$(0,1;0,95)$	**39,4%**	**39,4%**	40,6%	43,1%

Les résultats de la table 1 montrent principalement l'importance des facteurs d'échelle dans l'amélioration du taux EER pour cette mesure.
Ensuite, nous testons l'influence de l'introduction des probabilités bigrammes avec les mots précédents (Eq. 2). La mesure reste locale et ne dépend que du voisinage passé du mot hypothèse courant.

TAB. 2: Taux d'EER avec la probabilité bigramme arrière (Eq. 2).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	39,7%	39,9%	42,7%	44,7%
$(0,1;0,95)$	38,8%	**38,7%**	39,6%	42,8%

L'introduction des probabilités bigrammes se traduit par une amélioration des résultats (table 2), mais pas de manière importante. C'est pourquoi nous définissons une troisième mesure qui prend en compte un plus grand voisinage du mot hypothèse (Eq. 4). En effet, cette mesure se base sur les probabilités bigrammes avec les mots précédents et suivants. Là encore, nous observons une amélioration (table 3). Moyennant un retard de quelques trames, cette mesure peut encore être évaluée au cours de

la phase de construction du graphe de mots. Ce délai est nécessaire à l'obtention d'une stabilité dans le graphe pour la sélection des mots suivants.

TAB. 3: Taux d'EER avec les probabilités bigrammes avant et arrière (Eq. 4).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	42,5%	40,8%	39,7%	39,8%
$(0,1;0,95)$	39,8%	37,7%	**36,8%**	39,3%

La table 4 présente les résultats de notre dernière mesure de confiance à caractère local fondée sur la probabilité *a posteriori* locale et utilisant un voisinage plus important. Les résultats sont présentés selon la taille en mots de la *pseudo phrase* centrée sur le mot hypothèse courant. Pour déterminer la taille du voisinage en nombre de mots, nous nous basons sur la meilleure hypothèse de phrase. Avec un faible retard par rapport à la progression du moteur de reconnaissance, nous pouvons déterminer les mots qui précèdent et suivent le mot analysé.

TAB. 4: Taux d'EER avec la probabilité *a posteriori* locale.

		Taux de relâchement	
$(\alpha;\beta)$	Nb. mots	0,2	0,5
$(1;1)$	1	40,7%	40,8%
	3	37,1%	37,1%
	5	36,0%	36,0%
	7	35,6%	35,7%
$(0,1;0,95)$	1	43,4%	43,4%
	3	31,4%	31,5%
	5	**25,6%**	**25,6%**
	7	**24,3%**	**24,3%**

Nous pouvons remarquer l'influence de la longueur de la pseudo phrase. Plus on se rapproche de la longueur de la phrase complète, plus les résultats s'améliorent. Nous avons choisi comme mesure de référence la mesure de F. Wessel et al. [10] : mesure de la probabilité *a posteriori* calculée sur toute la phrase. A partir d'une pseudo phrase de 5 mots, la mesure locale donne des résultats proches de ceux de la mesure de référence (table 5). L'influence du voisinage diminue au delà d'une pseudo phrase de 5 mots, la longueur moyenne des phrases étant de 11,5 mots. Ce phénomène est sans doute dû à l'utilisation d'un modèle de langage bigramme. Ce phénomène reflète cependant l'aspect réel de la construction classique d'une phrase. Nous avons également testé cette mesure en définissant la taille du voisinage en nombre de trames et en extrayant le sous-graphe correspondant. Pour une taille de voisinage de 84 trames (longueur moyenne de 2 mots) de part et d'autre du mot analysé, nous obtenons le même résultat.

TAB. 5: Taux d'EER avec la probabilité *a posteriori* globale.

	Taux de relâchement	
$(\alpha;\beta)$	0,2	0,5
$(1;1)$	35,2%	35,1%
$(0,1;0,95)$	25,4%	**23,8%**

5. CONCLUSION

Dans cet article, nous avons présenté plusieurs mesures de confiance qui répondent à une contrainte forte : la mesure doit être utilisable pendant le processus de décodage du moteur de reconnaissance. Cela signifie que ces mesures ne nécessitent pas l'exécution complète du processus de reconnaissance pour être calculées. Au pire, certaines nécessitent un léger délai par rapport à la progression du moteur. Plus la mesure prend en compte d'information sur son voisinage proche, plus elle est pertinente. Une pseudo phrase de 5 mots est un bon compromis entre taux d'EER et délai pour pouvoir effectuer le calcul de la mesure. Notre meilleur taux d'EER, 24,3%, est atteint avec la mesure de la probabilité *a posteriori* locale et une pseudo phrase de 5 mots. Avec cette mesure, nous respectons notre objectif de réaliser une mesure trame-synchrone. Le taux d'EER obtenu est proche de celui de la mesure de référence, 23,8%, de F. Wessel et al. [10]. Ainsi, nous avons proposé des mesures de confiance pouvant être utilisées directement au cours du processus de reconnaissance ou bien encore à la demande pour valider un mot. Une continuation logique de ce travail consisterait à modifier automatiquement le processus de reconnaissance afin d'améliorer le taux du système.

RÉFÉRENCES

[1] P.R. Clarckson and R. Resenfeld. Statistical language modelling using the CMU-Cambridge toolkit. In *Eurospeech, Rhodes*, pages 2707–2710, 1997.

[2] L. Ferrer and C. Estienne. Improving performance of a keyword spotting system by using a new confidence measure. In *Eurospeech, Aalborg*, pages 2561–2564, 2001.

[3] T. Jitsuhiro, S. Takahashi, and K. Aikawa. Rejection of out-of-vocabulary words using phoneme confidence likelihood. In *ICASSP, Seattle*, pages 217–220, 1998.

[4] S.O. Kamppari and T.J. Hazen. Word and phone level acoustic confidence scoring. In *ICASSP, Istanbul*, 2000.

[5] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. In *Eurospeech, Aalborg*, pages 1691–1694, 2001.

[6] S. Ortmanns and H. Ney. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language*, 11 :43–72, 1997.

[7] J. Razik, O. Mella, D. Fohr, and J.P. Haton. Local word confidence measure using word graph and n-best list. In *INTERSPEECH, Lisbon*, pages 3369–3372, 2005.

[8] F. Wallhoff, D. Willett, and G. Rigoll. Frame-dicriminative and confidence-driven adaptation for LVCSR. In *ICASSP, Istanbul*, pages 1835–1838, 2000.

[9] M. Weintraub. LVCSR log-likelihood ratio scoring for keyword spotting. In *ICASSP, Detroit*, pages 297–300, 1995.

[10] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Proc.*, 9 :288–298, 2001.

Reconnaissance robuste de parole en environnement réel à l'aide d'un réseau de microphones à formation de voie adaptative basée sur un critère des N-best Vraisemblances Maximales

L. Brayda[1,2]*, C. Wellekens*[1]*, M. Omologo*[2]

[1]Institut Eurecom
2229 Route des Cretes , 06904 Sophia Antipolis, France
Mél : brayda,welleken@eurecom.fr - http ://www.eurecom.fr/ brayda

[2]ITC-irst
Via Sommarive 18, 38050 Povo (TN), Italy
omologo@itc.it

ABSTRACT

Distant-talking speech recognition in noisy environments is generally tackled by using a microphone array and a related multi-channel processing. Based on that framework, this paper proposes an *N-best* extension of the Limabeam algorithm, that is an adaptive maximum likelihood beamformer. *N-best* hypothesized transcriptions are generated at a first recognition step and then optimized independently one to each other. As a result, the *N-best* list is re-ranked, which allows selection of the maximally likely transcription to clean speech models. Results on real data show improvements over both Delay and Sum Beamforming and Unsupervised Limabeam at low SNR and with moderate reverberation.

1. INTRODUCTION

Les perfomances des systèmes de reconnaissance de la parole baissent nettement dans un environnement réel et d'autant plus que le locuteur se trouve loin du microphone dans un bruit soit additif, soit convolutif. Des études précédentes [1] ont montré que la qualité du signal vocal peut être améliorée (augmentation du rapport signal à bruit) par l'utilisation des réseaux de microphones. En exploitant la corrélation spatiale entre les signaux multi-canaux, on peut focaliser le réseau vers le locuteur (formation de voie ou *beamforming*). Ceci peut se faire soit en exploitant simplement l'interférence destructive du bruit par une technique de retard-et-sommation (R&S) [2], soit en appliquant des filtres à chaque canal (filtrage-et-sommation). Ces filtres peuvent être fixes ou adaptés pour chaque échantillon ou trame selon un certain critère [3, 4]. Le problème qui peut se poser est que l'amélioration du rapport signal à bruit (RSB) n'entraine pas nécessairement une augmentation concommittante de la performance du reconnaisseur [5]. Seltzer [6, 7] propose d' appliquer une technique de filtrage adaptatif sur les signaux multicanaux sous un critère de Vraisemblance Maximale (Limabeam) et non plus de rapport signal à bruit. Dans cette méthode les filtres sont adaptés de façon aveugle en utilisant les paramètres des modèles acoustiques non-bruités qui alignent le mieux les vecteurs acoustiques bruités. Le reconnaisseur utilisera ensuite la somme des signaux filtrés pour générer une transcription finale. Dans une étude récente [8] nous avons montré que dans un environnement simulé si l'on considère en parallèle les *N-best* hypothèses au lieu de la meilleure hypothèse pour adapter les filtres, on peut augmenter les performances du reconnaisseur et approcher celles d'un algorithme supervisé. Dans ce papier nous testons cette méthode améliorée dans un environnement réel et nous montrons que les performances du Limabeam peuvent être encore augmentées.

2. L'ALGORITHME LIMABEAM

L'algorithme Limabeam utilise un réseau de L microphones auquel on applique une formation de voie de type filtrage et sommation. Les coefficients des filtres non récursifs (RIF) de degré M, un par microphone, sont modifiés de façon adaptative. Un tel formateur de voie peut être représenté comme suit :

$$x[k] = \sum_{m=1}^{M} h_m[k] * s_m[k] \tag{1}$$

où $s_m[k]$ est le signal discret dans le domaine temporel reçu au m-ème microphone, $h_m[k]$ est la réponse impulsionnelle du filtre RIF du m-ème canal, $x[k]$ est la sortie du formateur, * dénote la convolution et k est l'index temporel. L'ensemble des filtres peut être représenté par un super-vecteur $\mathbf{h}$. Pour chaque trame, les composantes du vecteur acoustique sont calculées et exprimées en fonction de $\mathbf{h}$:

$$\mathbf{y}_L(\mathbf{h}) = \log_{10}\left(W\,|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))|^2\right) \tag{2}$$

où $\mathbf{x}(\mathbf{h})$ est le vecteur observé, $|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))|^2$ est le vecteur des différents composants du spectre de puissance, W est la matrice des coefficients Mel et $\mathbf{y}_L(\mathbf{h})$ est le vecteur des log-énergies des bancs de filtre (LFBE). Les coefficients cepstraux sont dérivés par une transformation en cosinus discrète (DCT).

$$\mathbf{y}_C(\mathbf{h}) = \mathrm{DCT}\left(\mathbf{y}_L(\mathbf{h})\right). \tag{3}$$

Limabeam vise à dériver un ensemble de *M* filtres RIF, qui maximisent la vraisemblance de $\mathbf{y}_L(\mathbf{h})$ étant donné un alignement Viterbi estimé d'une transcription supposée. Ceci est exprimé par :

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|w) \tag{4}$$

où w est la transcription supposée. L'optimisation est faite par gradient conjugué non linéaire. L'alignement Viterbi fait sur la sortie du réseau peut être estimé soit à partir de la transcription obtenue après une première étape de reconnaissance (Unsupervised Limabeam), soit en supposant que la transcription correcte est disponible (Oracle Limabeam). Plus de détails peuvent être trouvés dans [5]. L'Unsupervised Limabeam fonctionne bien dans les environnements bruyants, même avec un seul canal. Cependant, les expériences préliminaires conduites sur des données simulées [8] ont indiqué deux faits : d'abord, les résultats de l' Oracle Limabeam sur un canal simple étaient proches du R&S simple sur huit canaux ; en second lieu, il y avait toujours une marge d'amélioration possible entre l'Unsupervised et la version Oracle appliqués aux signaux multi-canaux.

3. APPROCHE *N-best* À OPTIMISATION PARALLÈLE

L'algorithme de Limabeam augmente la vraisemblance de l'hypothèse de la première transcription après une première étape de reconnaissance. Nous proposons d'appliquer *N-best* optimisations indépendantes et en parallèle : cette approche est basée sur le fait que la liste des *N-best*, avant l'optimisation parallèle, est triée par vraisemblance et pas nécessairement par le taux d'erreur en mots (WER), qui devrait être le critère optimal. En appliquant l'algorithme Limabeam sur chaque hypothèse, le tri de la liste des *N-best* hypothèses change parce que les hypothèses à WER inférieurs sont mieux optimisées, même s' ils ont une vraisemblance initiale inférieure. Nous prouvons au niveau expérimental que la nouvelle hypothèse choisie (la nouvelle hypothèse à vraisemblance maximale) dans cette nouvelle liste a, en moyenne, un WER inférieur à la première choisie dans la liste précédente (voir Figure 1)

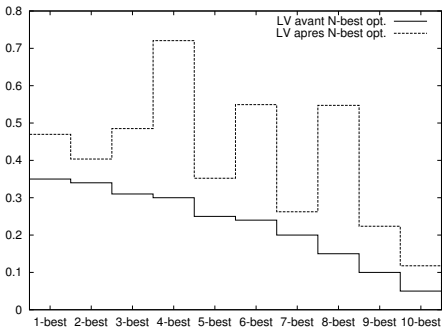


FIG. 1: *Exemple de Log-vraisemblance normalisée d'une phrase dont les 10 meilleures hypothèses sont optimisées. Avant l'optimisation les phrases sont triées par vraisemblance. Après, les vraisemblances de toutes les hypothèses sont augmentées et l'hypothèse 4, qui a un WER inférieur à l'hypothèse 1, est maintenant la meilleure.*

Il est à noter qu'ici "*N-best*" résulte d'une réduction préliminaire à une liste qui n'inclut pas de répétitions de la même phrase, qui pourraient résulter d'un nombre et de localisations différentes d'unités de bruit ou de silence. Le système est décrit ci-dessous. Pour chacune des *N-best* hypothèses nous dérivons un ensemble de filtres RIF :

$$\hat{\mathbf{h}}_n = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|w_n\,) \tag{5}$$

où w_n est la transcription à la première étape de reconnaissance, $P(\mathbf{y}(\mathbf{h})\,|w_n\,)$ est la vraisemblance de la phrase observée étant donnée la *n*-meilleure transcription considérée. Notez que l'équation (5) est équivalente à Unsupervised Limabeam quand *n* est 1. Après que tous les *N-best* vecteurs RIF aient été optimisés en parallèle, de nouveaux vecteurs acoustiques sont calculés et une deuxième étape de reconnaissance est exécutée. La transcription de vraisemblance maximale est alors choisie :

$$\hat{n} = \arg\max_{n} P(\mathbf{y}_C(\hat{\mathbf{h}}_n)\,|\hat{w}_n\,) \tag{6}$$

où $\hat{w}_n$ est la transcription produite à la deuxième étape de reconnaissance et $\hat{n}$ est l'index de la transcription la plus vraisemblable, soit $\hat{w}_{\hat{n}}$. L'optimisation est faite dans le domaine LFBE, alors que la reconnaissance est faite dans le

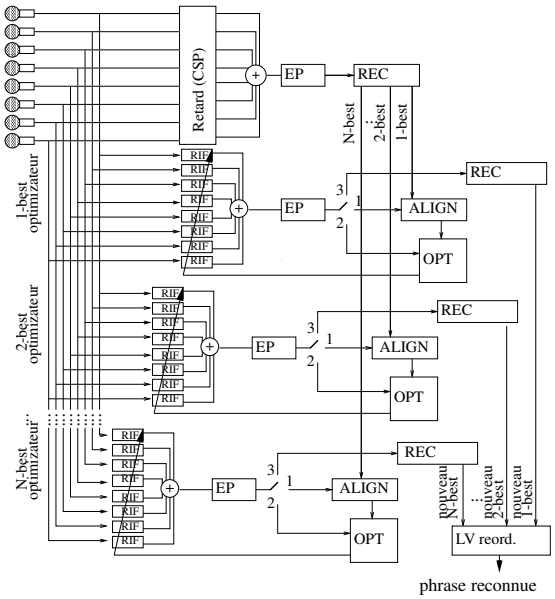


FIG. 2: *Schema du N-best Unsupervised Limabeam.*

domaine Cepstral comme dans [7]. Le nouvel ordonnancement des vraisemblances est de même fait dans le domaine Cepstral. Le système que nous proposons est représenté à la Figure 2. Le signal venant d'un réseau de microphones est traité par l'intermédiaire d'un R&S conventionnel, puis l'extraction des paramètres acoustiques (EP) et une première étape de reconnaissance est exécutée (REC). Le système de reconnaissance basé sur les modèles de Markov cachés (HMM) produit *N-best* hypothèses. Pour chaque hypothèse et en parallèle, l'algorithme de Limabeam est appliqué : d'abord un alignement Viterbi est effectué (commutateur sur 1 : ALIGN) et fixé, puis les coefficients des filtres RIF sont optimisés de manière adaptative en appliquant un algorithme de gradient conjugué (commutateur sur 2 : OPT). Un fois que la convergence est atteinte, les *N-best* séquences de vecteurs acoustiques sont identifiées (commutateur sur 3 : REC) et un ensemble différent de nouvelles transcriptions est produit. En conclusion, le dernier bloc compare les nouvelles *N-best* Log-Vraisemblances (LV-réordonnancement) en choisissant la plus élevée et la phrase reconnue est produite. Nos expériences montrent qu'en appliquant une approche N-best, l'Oracle Limabeam tel que proposé dans [7] ne constitue plus une limite supérieure à la performance du Limabeam : un alignement du type Baum Welch devrait produire une correspondance plus fine, avec en conséquence une meilleure optimisation. Pour obtenir une nouvelle borne, nous avons introduit la connaissance de la phrase correcte dans le bloc LV-reordonnacement : au lieu de (6), on choisit l'hypothèse dont la distance à la phrase connue est minimale. L'approche en aveugle *N-best* est donc couplée à une évaluation *a-posteriori* de la meilleure hypothèse : ceci est un indice de la qualité de la vraisemblance comme critère de choix.

4. BASE D'ÉVALUATION ET ENVIRONNEMENTS

Les expériences ont été conduites à l'aide du système de reconnaissance HTK basé sur les HMM, entraîné sur le corpus TI-digits. Les modèles de mots sont représentés

par des HMMs de type gauche-droite à 18 états, dont les distributions sont définies par une Gaussienne. La base d' éntrainement se compose de 8440 phrases, prononcées par 110 locuteurs (55 hommes et 55 femmes). La base de test se compose de 1001 phrases : elle a été enregistrée dans la salle décrite en Figure 3.

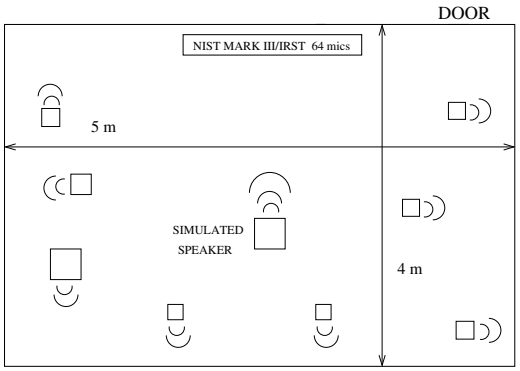


FIG. 3: *Chambre d'acquisition des données : les données non-bruitées sont émises par le haut-parleur central, le bruit simultanément par 8 sources. Le RSB à la source est de 0dB.*

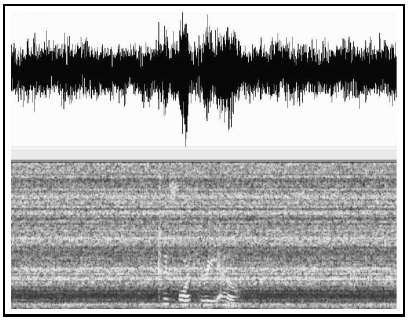


FIG. 4: *Spectrogramme d'une phrase enregistré avec un microphopne du MarkIII : le microphone capte les 8 bruits émis par les sources simultanées et distribuées et la phrase non-bruitée émise par le haut-parleur face au réseau.*

La salle mesure 5 x 4 mètres et elle présente un temps de propagation court (143 ms), ce qui nous a permis dans un environnement réel d'étudier les effets du bruit additif plutôt que ceux du bruit convolutif. Les données ont été produites par un haut-parleur de haute qualité (Tannoy 600A Nearfield Monitor). Le bruit, dont le spectre est visible dans la figure 4 a été engendré par 8 sources simultanées pour un rapport signal-bruit moyen de 0dB. Ce rapport est mesuré à la source et le vrai RSB mesuré au microphone varie selon la localisation des haut-parleurs et des microphones : on simule ainsi de meilleure façon un environnement réel. Les signaux ont été enregistrés par le réseau de microphones NIST MarkIII/IRST[9], placé à 1.3 mètres du haut-parleur Tannoy. Le MarkIII est un réseau linéaire de 64 microphones, dont les capteurs sont espacés de 2 cm. Pour nos expériences nous avons choisi d' utiliser 8 microphones, espacés de 16 cm : cette configuration représente un bon compromis entre les hautes performances qui dépendent du nombre de microphones, le respect du théorème du recouvrement spatial, le besoin d'une complexité gérable et d'un temps de réponse raisonnable (pour l'optimisation des filtres).

Le MarkIII acquiert les données audio à une fréquence d'échantillonage de 44.1 kHz : dans cet environnement reel, nous avons constaté que les performances dépendent relativement peu de la fréquence d'échantillonage, et donc les données ont été sous-échantillonnées à 16 kHz avec un filtre polyphase à trois étapes. Les filtres RIF ont une longueur de 10 échantillons. L'extraction des paramètres acoustiques génère 12 coefficients Cepstraux en échelle Mel (MFCC) et la log-énergie ainsi que les premières et deuxièmes dérivées, pour un total de 39 coefficients. Les paramètres ont été calculés chaque 10 ms, en utilisant une fenêtre glissante de Hamming de 25 ms. La gamme de fréquences couverte par le banc de filtres a été limitée à 100-7500 Hertz pour éviter des bandes de fréquence où l'énergie de parole est limitée. La normalisation par la moyenne du cepstre est appliquée (CMN). Alors que la reconnaissance est exécutée dans le domaine cepstral, l'optimisation est faite dans le domaine LFBE en utilisant des vecteurs acoustiques d'ordre 16 et une distribution Gaussienne pour les modèles, mais sans CMN [7]. L'implémentation de l'algorithme Limabeam n'a nulle part été modifiée par rapport au travail original afin d'assurer la conformité aux expériences de Seltzer.

5. RÉSULTATS ET DISCUSSION

L'environnement choisi permet d'obtenir a-priori des hautes performances avec une technique R&S, qui marche d'autant mieux que le bruit additif est diffus. Ceci est évident en regardant les performances de chaque microphone (Tableau 1) et celles du R&S (première ligne du Tableau 2) : les microphones plus proches du haut-parleur

TAB. 1: *Performance du reconnaisseur sur chaque microphone choisi du MarkIII. Les meilleurs résultats sont observés là où le microphone est le plus proche du haut-parleur. Résultats fournis en précision de mots, c'est à dire en tenant compte des insertions.*

mic	1	9	17	25
Pre.	50.76%	57.26%	63.91%	61.46%
mic	33	41	49	57
Pre.	62.52%	64.21%	62.76%	52.69%

TAB. 2: *Performance des différents formateurs de voie : R&S, Unsupervised Limabeam (U.L.), N-best Limabeam (N-best L.), Oracle Limabeam (O.L.) et a-posteriori N-best Limabeam (a-post). L'optimisation considère jusqu'à 40 hypothèses en parallèle. Pour chaque méthode, on indique si l'optimisation est aveugle (AV) ou supervisée (SUP), son résultat en précision de mots (Pre) et son amélioration relative (AR) par rapport au R&S. Il est à noter que le a-posteriori N-best est une limite supérieure de reconnaissance par l'N-best Limabeam, parce qu'il optimise les RIF de façon aveugle, mais choisit, de façon supervisée, la phrase qui maximise la précision au lieu de celle qui maximise la vraisemblance.*

Mèthode	SUP	AV	Pre	AR
R&S	-	-	80.74%	-
U.L.		X	83.16%	12.5%
O.L.	X		83.49%	14.2%
N-best L. (40).		X	83.83%	16%
a-post (40)	X	X	85.13%	22.8%

ont les meilleures performances et l'absence d'une symétrie des résultats par rapport au centre du réseau (microphone 33) est une conséquence de la diffusion non-symétrique du son et du bruit additif et convolutif dans la salle. Pour appliquer le R&S, les retards appliqués aux signaux multi-canaux sont estimés avec l' information de la phase du spectre croisé (CSP) [10, 11]. La bonne performance du R&S (80.74%) est atteinte grâce à l' échantillonage spatial efficace du réseau.

La Figure 5 montre le comportement du *N-best* Limabeam en fonction de la longueur de la liste des *N-best* hypothèses. Le point de départ de la courbe (83.16%) correspond au Unsupervised Limabeam, quand une seule hypothèse est considérée. Les résultats s'améliorent d'autant plus que la liste est longue. Un phénomène apparemment surprenant est le fait que le *N-best* Limabeam se situe au dessus de l'Oracle Limabeam : comme prévu en Section 2, un alignement qui considère tous les chemins pourrait augmenter les performance de l'Oracle. La courbe semble présenter une asymptote au delà des 34-meilleures hypothèses et y atteint le maximum de 83.86%. Ceci est dû à la présence de bonnes hypothèses dans la partie inférieure de la liste *N-best* et indique que considérer plus d'hypothèses est la clé pour obtenir de meilleurs résultats. Pour des RSB plus élevés, l'asymptote des performances devrait être atteinte plus vite, c'est à dire en considérant moins d'hypothèses. Le comportement non-monotone, visible aussi dans les résultats rapportés en [8], est dû à l'inconsistance entre les critères du maximum de vraisemblance et du minimum WER parce que nous savons que choisir la transcription la plus vraisemblable dans le bloc d'ordonnancement (cfr. Figure 2) n'implique pas un choix du type WER minimum. Ceci n'est pas le cas si l'on observe le comportement *a-posteriori N-best* Limabeam, où la courbe est strictement monotone. Ceci car considérer plus d'hypothèses accroit nécessairement les chances de choisir l'hypothèse correcte lorsque la décision repose sur un critère WER qui ne peut pas s'appliquer car on ne connaît pas la phrase correcte d'avance. Les améliorations absolues et relatives sont rapportées à la Table 2 : l'utilisation de Limabeam est clairement justifiée et dans cet environnement, les performances de la méthode non-supervisée se rapprochent de l'Oracle. Comme on peut l'observer à la Figure 5, une approche *N-best* dépasse l'Oracle fournissant une amélioration relative de 16% par rapport au R&S lorque 40 hypotheses sont traitées en parallèle. Dans les mêmes conditions, le *N-best* Limabeam *a-posteriori* peut atteindre une amélioration relative de 22.8% : ceci signifie qu'en modifiant le critère de réordonnancement, on peut atteindre les performances d'un reconnaisseur *a-posteriori*. Une façon de réaliser cet objectif est de pondérer d'avantage les hypothèses dont le LV croît plus pendant le pas d'optimisation. Cette solution est encore à l'étude. En outre, au cours de ce travail nous avons amelioré les taux de reconnaissance dans un environnement de bruit diffus dans lequel une formation de voie adaptative apporterait un gain généralement inférieur par rapport au R&S que dans des environnements à bruit plus directif. Ceci nous encourage à explorer différents environnements bruités et réverbérents c'est à dire de nous rapprocher des conditions typiques d'une salle de réunion.

6. REMERCIEMENTS

La collection des données à eté partiellement supportée par le projet de recherche IST EU FP6 HIWIRE. L. Brayda voudrait remercier le MESR (Ministère de l'Enseignement Supérieur et de la Recherche - France) et Istituto Trentino di Cultura pour avoir supporté ce travail.

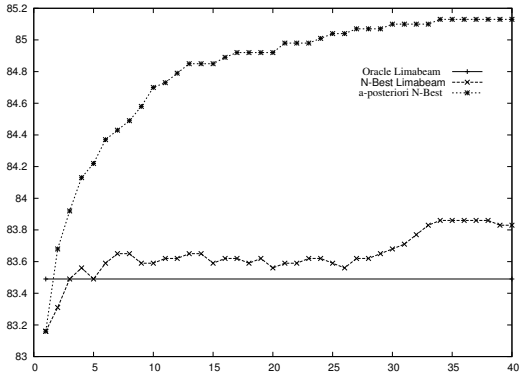


FIG. 5: *Performance de l' Oracle, du N-best Limabeam et du a-posteriori N-best Limabeam en fonction du nombre d'hypothèses considerées. Resultats exprimés en précision.*

RÉFÉRENCES

[1] M. Brandstein and D. Ward, *Microphone arrays - signal processing techniques and applications*, New York : Springer-Verlag, 2001.

[2] Johnson D and D. Dudgeon, *Array signal processing*, Prentice Hall, 1993.

[3] L. Griffith and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," in *IEEE Trans. on Antennas and Propagation*, 1982, vol. AP-30, pp. 27–34.

[4] O. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, 1972, vol. 60, pp. 926–935.

[5] Seltzer M., *Microphone array processing for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.

[6] Seltzer M. and Raj B., "Speech recognizer-based filter optimization for microphone array processing," in *IEEE Signal Processing Letters*, March 2003, vol. 10, no, 3, pp. 69–71.

[7] Seltzer M., Raj B., and Stern R. M., "Likelihood-maximizing beamforming for robust hands-free speech recognition," in *IEEE Trans. on Speech and Audio Procesing*, September 2004, vol. 12, no, 5, pp. 489–498.

[8] Brayda L., Wellekens C., and Omologo M., "N-best parallel maximum likelihood beamformers for robust speech recognition," in *accepted to Proceedings of EUSIPCO*, Florence, Italy, 2006.

[9] Brayda L., Bertotti C., Cristoforetti L., Omologo M., and Svaizer P., "Modifications on NIST MarkIII array to improve coherence properties among input signals," in *AES, 118th Audio Engineering Society Convention, Barcelona, Spain*, 2005.

[10] M. Omologo and P. Svaizer, "Acoustic event localization using a cross-power spectrum phase based technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[11] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, vol. 24, no, 4, pp. 320–327.

Les Nasales du Portugais et du Français : une étude comparative sur les données EMMA

Solange Rossato[1], António Teixeira[2], Liliana Ferreira[2]

[1]Institut de la Communication Parlée, CNRS/INPG/Université Stendhal,
BP 25 38040 Grenoble Cedex 9, France
Solange.Rossato@icp.inpg.fr
[2]Dep. Electrónica e Telecomunicações/IEETA, Universidade de Aveiro
Campus de Santiago, 3810 193 AVEIRO, Portugal
lsferreira@ieeta.pt, ajst@det.ua.pt

ABSTRACT

In this paper we present a first comparative study of velum height and movement in French and Portuguese based on EMMA data. Results show that the velum height reaches the highest position for oral consonants and decreases for oral vowels, nasal consonants and nasal vowels for both languages. Open vowels were found to be pronounced with velum height similar to the height used in nasal consonants production. Nasal vowels are produced with the lowest velum height in both languages but reveal different dynamic patterns.

1. INTRODUCTION

Le trait de nasalité est présent dans 97% des langues de la base UPSID en ce qui concerne les consonnes tandis que seulement 20% de ces langues possèdent des voyelles nasales. Le français et le portugais exploitent tous deux le trait de nasalité dans leurs systèmes vocalique et consonantique. En portugais européen, il y a cinq voyelles nasales, plusieurs diphtongues nasales et quelques triphtongues. Les voyelles nasales sont /ĩ ẽ ɐ̃ õ ũ/ que l'on trouve dans des mots tels que *sim* [sĩ] « oui », *penso* [pẽso] « je pense », *lã* [lɐ̃] « laine », *com* [kõ] « avec », *um* [ũ] « un ». Le processus de nasalisation des voyelles en portugais_résulte, dans la majorité des cas, de l'assimilation régressive depuis la consonne nasale postposée [1]. C'est d'un même processus d'assimilation qu'émergent les voyelles nasales du français [ɛ̃ œ̃ ɔ̃ ɑ̃].

Le velum, articulateur principal de la nasalité : Le trait de nasalité est réalisé grâce à l'abaissement du velum permettant la connexion des fosses nasales au conduit oral. Cependant, une ouverture vélopharyngée n'implique pas forcément un son perçu comme nasal. Le velum peut être abaissé lorsqu'un locuteur produit la voyelle orale [a], et ceci même en dehors de tout contexte nasal, ainsi que le souligne Durand [2] (p 34), d'après les radiographies de Clumsky, pour une réalisation du [a] de *il l'a* avec un velum distant de 10 mm de la paroi pharyngale. Notons cependant que ce phénomène n'a pas été retrouvé dans les travaux de Bothorel et al. [3] dans les contextes oraux étudiés.

Plusieurs études montrent que la présence d'un passage vélopharyngé n'est pas une condition suffisante à la perception de la nasalité. Maeda [4] utilise un modèle articulatoire pour synthétiser un continuum depuis la voyelle orale jusqu'à la voyelle nasale et montre qu'une faible ouverture vélopharyngée ne permet pas la perception de la nasalité pour la voyelle [a], tandis qu'elle suffit pour la voyelle [i]. De même, Warren et al. [5] note qu'une ouverture vélopharyngée supérieure à 0.2 cm² est nécessaire pour percevoir la nasalité. On ne peut donc pas décrire le contraste de nasalité en termes de fermeture et d'ouverture du port vélopharyngé mais de degré de couplage : le velum doit s'abaisser suffisamment pour produire ce trait de nasalité. Quelles sont les positions cibles que doit atteindre le velum pour réaliser ce contraste de nasalité ? Avec quelle précision doit-on atteindre ces cibles ? Ces cibles sont-elles les mêmes en français et en portugais européen ? Pour répondre à ces questions, nous allons comparer les positions cibles du velum chez un locuteur français et chez un locuteur portugais.

Dynamique du geste articulatoire : L'aspect dynamique des voyelles nasales du portugais est mentionnée par Lacerda et Stevens (1956) (cité par [1]) : « *According to a number of phonetic studies, the nasal vowels of Portuguese differ from the nasal vowels of, for example, French, in that they are strongly nasalized only near the end* ». Cet aspect dynamique des voyelles nasales est souligné par Ohala [6] qui observe qu'en Hindi, la voyelle nasale montre un abaissement progressif plus important durant sa production que ne le fait une voyelle nasalisée par contexte. D'autre part, les voyelles nasales du français sont interprétées par Feng et Castelli [7] comme une tendance depuis la configuration orale de la voyelle vers une configuration cible que constitue le conduit pharyngo-nasal. Clumeck [8] est un des premiers à proposer le rôle de la dynamique dans la perception de la nasalité. Après avoir étudié les hauteurs du velum chez des locuteurs Américains, Suédois, Amoy, Hindi et portugais Brésiliens, l'auteur conclut sa discussion en ces termes : *« It might then be the case that the listener's perception of the presence or absence of nasalization is more dependent upon the timing of palatal lowering rather than upon actual extent of*

palatal lowering ». Teixeira [9] a montré que la dynamique du velum et d'autres articulateurs permettait d'améliorer l'aspect naturel des voyelles nasales synthétisées à l'aide d'un synthétiseur articulatoire. C'est pourquoi nous nous intéressons également aux aspects dynamiques des voyelles nasales du portugais et du français.

2. CORPUS ET MESURES

Cette étude analyse deux corpus l'un portugais l'autre français présentés respectivement dans les études de Teixeira [9] et Rossato et al. [10]. L'objectif est ici d'uniformiser les analyses pour comparer les mouvements du velum chez nos deux locuteurs.

2.1. Corpus EMMA

Corpus portugais : Le corpus a été construit pour caractériser les mouvements du velum lors de la production des voyelles nasales, comparer avec les voyelles orales, déterminer les variations de la position du velum dans les structures de type $C_1V_nC_2$ où V_n est une voyelle nasale et C_1 et C_2 sont deux plosives, et lorsque la voyelle nasale suit une consonne nasale. Une deuxième partie est constituée de différents contextes (fricatives, latéral, vibrante) et de phrases, partiellement exploitée ici, notamment pour augmenter le nombre de voyelles orales. Les enregistrements ont eu lieu à Ludwigs Maximillians Universität, Munich, avec un articulographe Carstens AG100 de 10 pellets (seulement 9 ont été utilisés). Le sujet, un des auteurs, est un locuteur mâle de 32 ans de langue maternelle portugaise. Trois pellets ont été fixés sur la langue dans le plan médiosagittal, un sur la lèvre inférieure, deux pellets ont servi de référence. Le pellet mesurant les mouvements du velum a été fixé sur une languette en plastique prolongeant un palais artificiel. La totalité du corpus a pu être enregistré sans qu'aucun des pellets ne se décolle.

Corpus français : Le corpus français s'intéressait à la position du voile du palais et à ses variations lors de la réalisation des sons du français, notamment les oppositions orales/nasales. Pour cela le corpus est constitué de séquences VCV où V = /i y u e ø o ɛ œ ɔ a ɛ̃ œ̃ ɔ̃ ɑ̃/, et C = /p t k b d g f s ʃ v z ʒ m n ʁ l/, de séquences [pVCV] avec V = /i u a ɛ̃ œ̃ ɔ̃ ɑ̃/ et C = /p t b d m n /, répétées 3 fois, ainsi que de phrases et de nomogrammes. Les séquences VCV sont analysées ici. Les enregistrements ont eu lieu à l'Institut de la Communication Parlée, Grenoble avec un articulographe Carstens AG100 disposant de 5 pellets. Le sujet est un locuteur mâle de langue maternelle française. Deux pellets ont été collés sur les incisives supérieures et inférieures, deux sur la langue et un sur le velum. Les 5 pellets sont restés fixés durant toute la session d'enregistrement.

2.1. Mesures des mouvements du velum

Référence : Pour ces deux enregistrements EMMA, ce n'est ni le même appareil ni le même locuteur. Et même lors de deux enregistrements du même locuteur, on ne peut garantir de fixer le pellet au même endroit. Les données ne sont donc pas directement comparables. Ceci étant dit, ces deux enregistrements fournissent une indication des mouvements du velum. Nous avons mesuré la hauteur du velum VH (coordonnée Y du pellet du velum) et nous avons pris comme référence 0 la hauteur la plus haute observée dans chaque corpus. Cette position la plus haute est observée pour [k] pour le locuteur français, et pour [g] dans le corpus du portugais. Ainsi, VH varie entre 0 cm et -1,2 cm pour le velum le plus abaissé. Nous avons choisi de conserver les unités métriques sans normaliser. En effet, s'il est connu que les plosives sont réalisées avec un velum relevé, nécessaire à la fermeture du port vélopharyngé, et cela quelle que soit la langue, rien n'indique que la position la plus basse observée durant les voyelles nasales ne soit comparable entre les deux langues. Les différences observées peuvent être imputables à la technique de mesure, (position du pellet, film plastique ou directement sur la muqueuse…) mais également à une différence de cibles entre les deux langues.

Amplitude du geste : Nous avons extrait VH au milieu du phone, les transitions étant repérées manuellement à l'aide des signaux acoustiques et des sonagrammes. Pour cet étiquetage, aucune information concernant les trajectoires du velum n'est utilisée. Cette mesure a l'avantage de pouvoir être obtenue pour tous les types de sons, indépendamment du contexte. Cependant, cela ne correspond pas à une réalité articulatoire et l'abaissement maximal ne se situe pas forcément au milieu de la voyelle nasale. C'est pourquoi nous avons également détecté, à l'aide de la trajectoire du pellet du velum et de sa dérivée, la VH la plus basse de la trajectoire.

Mesures dynamiques : Pour analyser la dynamique du velum, nous avons mesuré les vitesses de variation de VH lors de la réalisation des voyelles nasales, plus spécifiquement lors de la phase d'abaissement d'une part, et lors de la phase de remontée du velum d'autre part. En portugais, lors des séquences CVC, les deux phases se retrouvent dans la même voyelle nasale tandis que dans les séquences VCV du corpus français, la phase de remontée est située dans la 1ère voyelle et la celle d'abaissement dans la 2ème voyelle.

3 GESTES ARTICULATOIRES DU VELUM

3.1 Hauteur du velum et contraste de nasalité

Le contraste de nasalité oppose les consonnes orales C et les consonnes nasales N ainsi que les voyelles orales V et les voyelles nasales VN. Dans un premier temps,

nous avons mesuré VH au centre du segment quel que soit le contexte, et ce sur les deux corpus portugais et français. Cependant, les voyelles orales sont nasalisées en contexte nasal, et donc réalisées avec un velum plus bas, nasalisation que l'on perçoit [11]. Ici, nous voulons opposer les cibles articulatoires des voyelles orales et des voyelles nasales, nous n'avons donc pas pris en compte les voyelles orales en contexte nasal. La figure 1 présente la répartition des VH mesurées sur chacun des deux corpus. Les cibles (valeurs moyennes de VH) montrent la même hiérarchie en français et en portugais : C > V > N > VN.

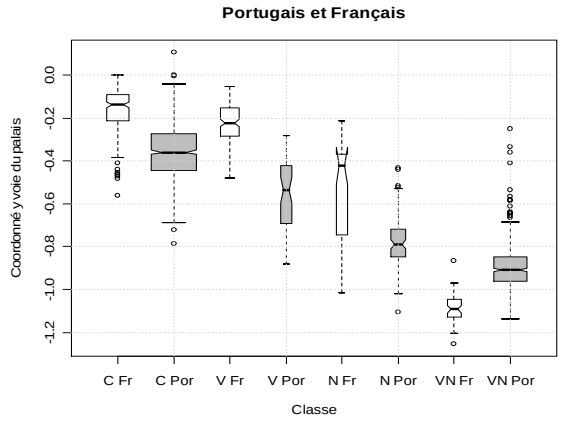


Figure 1 – Hauteurs du velum (en cm) pour chaque classe C, V, N et VN sur le corpus français (blanc) et portugais (gris).

Contraste de nasalité consonantique : Les consonnes orales sont les phonèmes réalisés avec la cible la plus haute et une latitude de réalisation relativement faible en français. La cible est légèrement plus basse, avec une tolérance plus grande pour les consonnes orales du portugais, dont le corpus est plus varié. Les consonnes nasales sont réalisées avec une cible articulatoire bien plus basse dans les deux langues. On note cependant un recouvrement des VH entre consonnes orales et nasales (entre -0.2 et -0.4 cm en français et -0.5 et -0.7 cm en portugais) qui laisse supposer que certaines consonnes orales sont produites avec un port vélopharyngé ouvert. On observe pour les consonnes nasales une latitude de variation importante autour de la cible notamment en français alors qu'elle ne concerne dans les deux corpus que les segments [m] et [n], soulignant que ces deux sons tolèrent une grande variabilité de la position du velum.

Contraste de nasalité vocalique : Les voyelles orales et nasales ont des cibles articulatoires très distinctes en français, ce qui n'est pas le cas en portugais où l'on a un recouvrement autour de VH = -0.8 cm. Les figures 2a) et 2b) détaillent les VH de certaines voyelles, ainsi que des consonnes nasales. La voyelle ouverte [a] est produite, avec une VH plus basse que celles des autres voyelles orales, assez proche des VH des consonnes nasales [m] et [n]. Il ne peut s'agir d'un effet de coarticulation puisque nous n'avons considéré que les voyelles en isolation ou en contexte oral. Il semble donc que la voyelle [a] soit produite avec un port vélopharyngé ouvert.

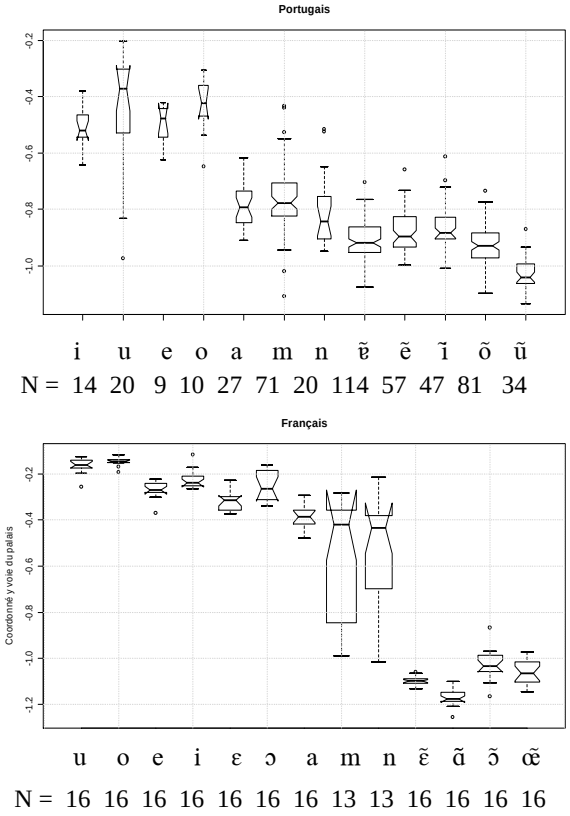


Figure 2 – VH par phonème (N est le nombre de segments analysés) a) en haut pour le portugais, b) en bas pour le français.

Les voyelles nasales du français ont des cibles articulatoires situées vers -1 à -1.2 cm d'abaissement et tolèrent très peu de variation tandis que les voyelles nasales du portugais ont des cibles autour de -0.9 cm qui montrent une variabilité plus importante. Nous avons alors mesuré l'abaissement maximal de la voyelle nasale, pour vérifier que ces différences n'étaient pas un artefact dû au fait que l'on prend VH au milieu du phonème, et les résultats des deux mesures sont similaires en français et en portugais (différence significative seulement pour le /ɐ̃/ portugais). On ne peut pour autant affirmer que la cible des voyelles nasales du français est plus basse que celle du portugais car les deux mesures ne sont pas exactement comparables.

3.2 Analyse des aspects dynamiques

Voyelles nasales en contexte oral : Nous avons mesuré la vitesse maximale de variation de la hauteur du velum, en cm/s, lorsque le velum s'abaisse et se relève pour chaque voyelle nasale en contexte de plosives (cf. Table 1). Les gestes d'abaissement et de

remontée du velum se font avec des vitesses similaires pour le sujet français (F=3.6, p=0.59), tandis que le geste d'abaissement est plus lent que celui de la remontée pour le sujet portugais (F=196, p<0.001). Cette différence peut être due au fait que les gestes du locuteur portugais sont analysés en CVC alors que ceux du locuteur français le sont sur des séquences VCV.

VN Fr	ɛ̃	œ̃	ɔ̃	ɑ̃
abais.	8.4 (*1.3*)	8.1 (*0.9*)	8.5 (*1.4*)	8.1 (*0.8*)
montée	8.9 (*1.6*)	7.9 (*1.2*)	7.2 (*1.4*)	7.0 (*1.9*)

VN Por	ĩ	ũ	ẽ	õ	ɐ̃
abais.	5.7(*1.1*)	7.7(*1.7*)	6.3(*1.8*)	7.4(*1.8*)	6.3(*1.8*)
montée	7.6(*1.8*)	10.2(*1.6*	7.8(*1.7*)	8.4(*1.6*)	8.9(*1.8*)

Table 1 – Vitesse de variation de VH pour les voyelles nasales du français (en haut) et du portugais (en bas). Les valeurs entre parenthèses sont les écart-types.

Voyelles nasales après une consonne nasale : Les figures 3 et 4 illustrent les trajectoires de VH en portugais et en français. On observe des transitions N - VN assez différentes. En portugais, le velum remonte lentement durant la consonne nasale pour atteindre son point le plus haut au début de la voyelle nasale, avant de redescendre et d'atteindre la position la plus basse de la voyelle. Ce mouvement de remontée durant le [m] n'est pas observé en français : partant de la cible très basse de la voyelle nasale précédente, le velum reste stable pendant la consonne, commençant à descendre peu avant le début de la voyelle pour rejoindre la position stable de la voyelle nasale.

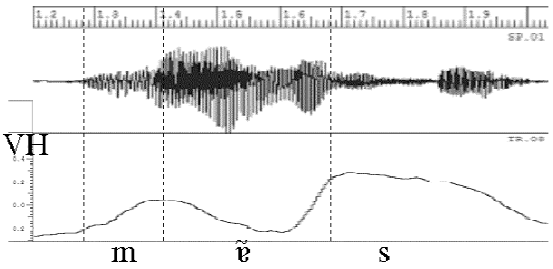


Figure 3 – Signal acoustique et trajectoire de VH en fonction du temps (s) pour /mɐ̃s/ en portugais.

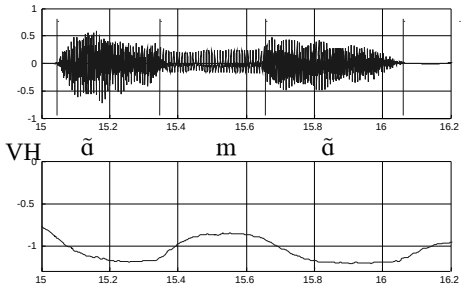


Figure 4 – Signal acoustique et trajectoire de VH en fonction du temps (s) pour /pɑ̃mɑ̃/ en français.

4. CONCLUSION

Dans nos deux corpus portugais et français, nous retrouvons la même hiérarchie dans les positions cibles C>V>N>VN, ainsi qu'une réalisation des voyelles orales ouvertes avec VH du même ordre que les consonnes nasales, impliquant vraisemblablement un port vélopharyngé ouvert. Les mesures de vitesses ont montré une différence entre la dynamique des voyelles nasales du portugais et celle du français qu'il faudrait étudier dans des contextes comparables. Nous n'avons pas abordé ici l'influence du contexte, et il serait intéressant de voir si les phénomènes de coarticulation sont similaires dans ces deux langues. Se pose aussi la question de l'importance de ces trajectoires dynamiques différentes dans la perception des voyelles nasales par des locuteurs français et portugais.

REMERCIEMENTS

Tous nos remerciements vont à nos deux sujets Pierre Badin et Antonio Teixeira. Merci à Phil Hoole et Christophe Savariaux pour le recueil des données respectivement du portugais et du français. La partie sur le portugais a été financée par le projet FCT POSC/PLP/57680/2004 HERON de l'Agence de Recherche Portugaise. La partie sur le français est soutenue par un financement ANR : « Dynamique de la nasalité ».

BIBLIOGRAPHIE

[1] R. Sampson. *Nasal Vowel Evolution in Romance*. Oxford University Press, 1999.

[2] M. Durand. De la formation des voyelles nasales In *Sudia Linguistica VII*, p33-53, 1954.

[3] A. Bothorel, P. Simon, F. Wioland & J.P. Zerling, *Cinéradiographie des voyelles et consonnes du français*, Trav. Institut Phon.. Strasbourg, 1986.

[4] S. Maeda. Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. In [12]. p147-167, 1993.

[5] D.W. Warren, T.M. Dalston, et al. Aerodynamics of Nasalization, In [12], p119-146, 1993.

[6] J. Ohala & M. Ohala. The phonetics of nasal phonology. In [12], p225-249, 1993.

[7] G.Feng & E. Castelli, Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization. *JASA*. 99(6) p3694-3706, 1996.

[8] H. Clumeck. Patterns of soft palate movements in six languages. *J. of Phonetics* 4, p337-351, 1976.

[9] A. Teixeira, F. Vaz, & J. C. Príncipe. Influence of dynamics in the perceived naturalness of Portuguese nasal vowels. In *Proc. ICPhS*, 1999.

[10] S. Rossato, P. Badin & F. Bouaouni. Velar movements in French: an articulatory and acoustical analysis of coarticulation. In *Proc. ICPhS*, Barcelona, p3141-3145, 2003.

[11] R.A. Krakow & P. S. Beddor. Coarticulation and the perception of nasality. *Proc. ICPhS*, 1991

[12] *Phonetics and Phonology, Vol 5, Nasals, Nasalization and the Velum*, San Diego, Academic Press, 1993

Une analyse prosodique de la parole souriante : étude préliminaire

Caroline Émond

Université du Québec à Montréal
Département de linguistique et de didactique des langues, C. P. 8888, Succ. Centre-ville, Montréal (Québec) H3C 3P8
caroemond@hotmail.com
www.phonetique.uqam.ca

ABSTRACT

Smile is a visible expression and it has been shown that it's audible too (Tartter [9], Schröder & al. [8], Aubergé & Cathiard [1]). The aim of this study is to describe the prosodic correlates of smiled speech produced by 6 speakers of Quebec French. In order to elicitate smile, subjects were required to read sentences containing caricatures. This condition was compared to a neutral condition, in which no caricatures were present. Utterances were digitized and analysed with Praat and F0 measurements were extracted. Results show that F0 register tends to be lower in normal speech than in smiled speech. Because the global shapes of F0 movements are similar in the two kinds of speech, we claim that this difference is phonetic rather than phonologic.

1. INTRODUCTION

Le rire, comportement humain remarquable, est présent dans toutes les cultures (Trouvain [12]) et il nous permet d'exprimer des émotions (associées habituellement à la joie et au plaisir). Selon Mowrer & al. [6], on définit généralement le rire comme étant à la fois une vocalisation et un comportement (vocal non verbal). Ce comportement est étrange, difficile à analyser et il a reçu beaucoup d'attention d'un point de vue psychologique, souvent en rapport avec l'humour. Les mêmes auteurs signalent aussi que le rire est un réflexe émotionnel instinctif, une réponse à un stimulus (humour, chatouillements, etc.) et qu'il est une mine d'or d'informations communiquées à travers les aspects visuels et auditifs. Toutefois, on en connaît peu sur les caractéristiques acoustiques du rire. Mais qu'en est-il du sourire exactement? S'agit-il d'une sorte de rire, mais de moindre intensité ou doit-on le considérer dans une classe à part? Comme plusieurs auteurs l'ont démontré, il est perceptible, mais là encore, tout comme pour le rire, on ne connaît à peu près rien de ses caractéristiques acoustiques. Or, un interlocuteur est tout de même capable d'identifier le rire et le sourire dans la voix de son vis-à-vis. S'il est possible de percevoir ces expressions, il est alors possible de leur trouver des caractéristiques communes, des paramètres acoustiques permettant de les identifier.

2. PROBLÉMATIQUE ET CADRE THÉORIQUE

2.1. Problématique

Ce qui ajoute à la particularité du rire c'est que, pour au moins un de ses aspects, il est paradoxal, en ce sens qu'il est à la fois stéréotypé et idiosyncrasique. Comme le souligne Provine [7], bien que « la plupart des gens rient […] de façon assez semblable » (p. 70), ils ne rient pas de façon identique. Pour Bachorowski & al. [2], dans leur imposante étude portant sur 1024 rires de 97 adultes, les paramètres acoustiques du rire sont trop variables et complexes pour qu'on puisse invoquer une stéréotypie du rire. Pour ces auteurs, les études antérieures ont accordé trop d'importance au caractère stéréotypé du rire et leurs résultats vont à l'encontre de cette assertion, car le rire apparaît plutôt comme un répertoire de sons ayant plusieurs sous-catégories. Nous sommes en accord à la fois avec Bacharowski & al. [2] et Provine [7]. Bien qu'il soit difficile d'analyser le rire en raison de sa complexité (Bacharowski & al. [2]), « si le rire ne comportait pas une certaine invariance, nous ne reconnaîtrions pas que les gens rient et le rire ne serait pas un signal social efficace » Provine [7] (p. 70). En fait, nous pensons que ce qui fait défaut dans le domaine du rire, c'est peut-être le manque d'une véritable typologie, une classification des genres qui faciliterait justement l'analyse de cette réalité complexe.

Le sourire, quant à lui, a reçu moins d'attention que le rire. Nous avons décidé d'appliquer au sourire ce que nous venons tout juste de décrire pour le rire (invariance vs stéréotypie et idiosyncrasie). Le sourire est une expression qu'il est possible de reconnaître visuellement et comme sa production implique un changement de la configuration du conduit vocal, par rapport à sa position neutre, il est fort probable dans ce cas, qu'il soit audible également (Tartter [9]). Il a été démontré qu'il est possible de percevoir le sourire du point de vue acoustique (Tartter [9], Schröder & al. [8]), mais son côté idiosyncrasique n'a pas permis jusqu'à maintenant d'identifier les paramètres précis et invariants qui lui sont associés (Aubergé & Cathiard [1]). Notre question de recherche est donc la suivante : Quels sont les corrélats prosodiques du sourire produits par des locuteurs et des locutrices du français québécois permettant de l'identifier? Notre objectif est de décrire ces corrélats prosodiques du sourire afin de

voir s'il est possible de dégager des points communs qui permettraient de mieux le définir d'un point de vue acoustique.

2.2. Cadre théorique

Dans plusieurs langues, on aurait tendance à considérer le sourire comme le « petit frère » du rire (Trouvain [12]), et de ce fait, à supposer qu'il pourrait se trouver sur le même continuum que ce dernier. Les commentaires et résultats des tests de perception de Trouvain [12] suggèrent en bout de ligne le rejet de l'hypothèse d'un continuum entre le rire et le sourire même s'il peut arriver parfois que ceux-ci partagent certaines propriétés acoustiques.

Comment alors faire la distinction entre les deux? Nous nous appuyons sur l'étude de Trouvain [13], car les termes utilisés ainsi que leurs définitions permettent de saisir les différences fondamentales entre rire et sourire. La principale distinction résiderait dans le fait que la parole peut être articulée simultanément au sourire, tandis que le rire est séparé de l'articulation. En d'autres mots, on peut parler en même temps que l'on sourit, alors qu'il est impossible de rire et de parler en même temps. On peut donc dire que le sourire est une forme du rire synchrone à la parole (« *speech-synchronous form of laughter* ») qui se divise en deux catégories : « *speech-laugh* » et « *smiled speech* » (Trouvain [13]). Nous traduisons ces termes respectivement par *parole rieuse* et *parole souriante*.

Dans la parole souriante, il est possible d'entendre, de percevoir le sourire, soit sur quelques syllabes seulement, soit dans sa totalité. Dans la parole rieuse, seulement deux syllabes, en moyenne, sont affectées par la production d'une sorte de rire qui ressemble plus à une aspiration ou à un trémolo durant le processus de phonation (Trouvain, communication personnelle). Même s'il est possible d'étoffer ces concepts, de leur apporter plus de nuance, nous nous en tenons, pour le moment, à ces brèves définitions qui sont suffisamment explicites et qui rendent compte de la distinction essentielle à faire entre les deux types de sourires dans la parole. Ici, nous nous intéressons à la parole souriante.

3. MÉTHODOLOGIE

3.1. Constitution et présentation du corpus

Nous avons d'abord sélectionné 10 caricatures réalisées par Chapleau [3], [4] parues en 2003 et en 2004 dans le quotidien La Presse. Le choix des caricatures, en ce qui a trait à leur côté humoristique, s'est fait de façon informelle par accord interjuges. La sélection s'est effectuée sur la base des critères suivants : Les évènements illustrés dans les caricatures devaient être encore assez évocateurs. Afin que les participants évitent de jouer le ou les personnages représentés dans les caricatures, celles-ci ne devaient pas avoir de phylactères. Les titres devaient avoir entre 12 et 24 syllabes, car même s'il est possible de percevoir le sourire sur une seule syllabe comme l'a démontré Tartter [9], nous nous appuyons sur Trouvain [12] qui suggère que le sourire est un évènement à long terme alors que le rire serait un évènement à court terme.

Suivant les mêmes critères, nous avons composé 20 autres énoncés : 10 énoncés seuls et 10 énoncés présentés avec des images de type neutre. Les 10 énoncés des caricatures étaient présentés sans le dessin et avec le dessin (total = 40 énoncés). Nous avons constitué 6 versions du corpus (pour les 6 participants) où les énoncés apparaissaient dans un ordre aléatoire, afin d'éviter qu'un énoncé ait un effet sur l'énoncé qui suivrait. Les 10 premiers énoncés, tout comme dans Schröder & al. [8], devaient tous être des énoncés « neutres », c'est-à-dire sans caricatures, pour que les participants ne se doutent pas, au début du moins, du but de l'expérimentation. Pour celle-ci, nous avons utilisé le logiciel Power Point.

Participants et enregistrements

Trois femmes et 3 hommes, recrutés en milieu universitaire, âgés entre 22 et 34 ans, ayant le français québécois comme langue maternelle, ont pris part aux enregistrements.

Les enregistrements ont eu lieu dans une chambre sourde où les participants ont été filmés et enregistrés, à l'aide d'un ordinateur portable IBM C17302, d'un caméscope numérique Panasonic AG-DVC30, d'une enregistreuse numérique (DAT) TASCAM et d'un microphone dynamique Shure Beta 58A.

Les participants étaient assis devant l'ordinateur portable, le micro était à environ 30 cm d'eux et ils étaient filmés à environ 45°. Les consignes apparaissaient à l'écran de l'ordinateur et nous ajoutions qu'ils devaient s'imaginer être dans un endroit agréable (chez un ami en train de prendre un verre par exemple) afin de leur faire oublier, autant que faire se peut, l'environnement dans lequel ils se trouvaient. C'est la raison pour laquelle nous étions présente durant l'expérimentation. Comme le rire (et le sourire selon nous) a un caractère interactif, nous voulions créer une atmosphère conviviale où les participants ne sentiraient pas l'absurdité de la situation s'ils devaient rire seuls dans la chambre sourde d'un laboratoire de phonétique. Nous ne nous empêchions donc pas de sourire lorsque les participants souriaient eux-mêmes. Une phase d'entraînement précédait les enregistrements. Juste avant, ils devaient lire à haute voix le formulaire de consentement qu'ils avaient signé. Ceci constituait l'échantillon de F0 de la parole normale, en situation de lecture. Nous n'avons pas pris les 10 premiers énoncés neutres, car, au début, la nervosité de certains participants aurait pu modifier la F0. Ceux-ci devaient lire les 40 énoncés et pour passer à l'énoncé suivant, ils appuyaient eux-mêmes sur la

touche du clavier correspondant à la barre d'espacement. L'enregistrement durait entre 5 et 10 minutes.

3.2. Test de perception

Nous avons numérisé les données avec Adobe Premiere et segmenté les énoncés avec Goldwave. Pour la sélection du sous-corpus, nous avons repéré acoustiquement les énoncés qui semblaient avoir été prononcés avec un sourire et nous avons visionné l'enregistrement afin de sélectionner tous les énoncés prononcés avec l'expression faciale (étirement des lèvres) correspondant au sourire. Au total, 32 énoncés ont été retenus.

Un accord interjuges (4 femmes et 2 hommes) a permis de déterminer le deuxième sous-corpus. Pour ce test de perception, nous avons ajouté 12 énoncés neutres aux 32 « souriants » déjà retenus (total = 44). Les personnes entendaient les énoncés une seule fois et devaient dire s'il s'agissait d'un énoncé produit avec un sourire dans la voix ou de façon neutre. Pour l'analyse, nous n'avons conservé que les séquences souriantes perçues de façon dominante par les 6 auditeurs, soit 4 réponses identiques (et plus) sur 6, pour un total de 12 énoncés.

3.3. Analyse acoustique

Nous avons analysé nos données avec Praat, un logiciel de traitement de la parole. Les données ont été segmentées en syllabes. Les mesures de F0 ont été extraites à l'aide de l'algorithme d'autocorrélation, par pas de 10 ms. Une moyenne de F0 a été déterminée, pour chaque phrase. De plus, toutes les séquences ont été transcrites à l'aide du système de transcription ToBI (voir Thibault [10]), qui implique l'assignation de tons haut (H) et bas (B) en différents points de la courbe intonative. Cette transcription nous permet d'étudier l'inventaire phonologique prosodique des séquences, alors que les mesures instrumentales de F0 correspondent à l'implémentation phonétique de cette suite de tons.

4. RÉSULTATS ET DISCUSSION

4.1. Inventaire phonologique : tons haut (H) et bas (B)

Comme les 12 énoncés de notre deuxième sous-corpus, n'avaient pas tous leur pendant neutre, nous en avons sélectionné 6 pour lesquels le problème ne se présentait pas, soit 1 par locuteur (total =12). Une analyse qualitative des tons H et B ne nous a pas permis de dégager un patron commun qui permettrait de différencier les énoncés neutres des énoncés de parole souriante.

4.2. Réalisation phonétique : moyenne et étendue de F0

Les 12 énoncés perçus comme étant souriants lors du test de perception (voir 3.2.) ont servi à cette partie de l'analyse. On peut voir aux figures 1 et 2 que la moyenne de F0 en parole normale est plus basse que la moyenne de F0 en parole souriante, sauf pour un énoncé chez les femmes et deux chez les hommes. Ce résultat est conforme à ce qu'ont rapporté Mowrer & al. [6] et Hirson [5], pour le rire et Tartter [9], pour le sourire. Pour ce qui est de l'étendue de F0, on constate, aux figures 3 et 4, qu'elle est plus grande en parole normale qu'en parole souriante, à l'exception de 2 énoncés chez les femmes.

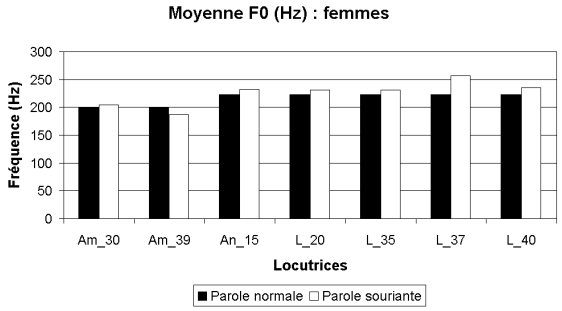


Figure 1 : Moyenne de F0 (Hz) pour 7 énoncés produits par 3 locutrices, condition neutre et sourire

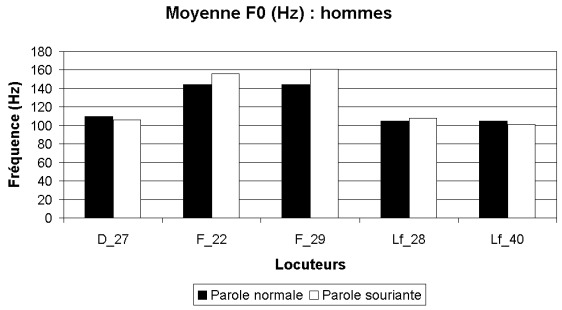


Figure 2 : Moyenne de F0 (Hz) pour 5 énoncés produits par 3 locuteurs, condition neutre et sourire

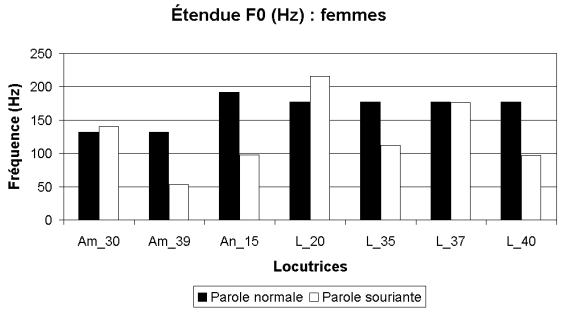


Figure 3 : Étendue de F0 (Hz) pour 7 énoncés produits par 3 locutrices, condition neutre et sourire

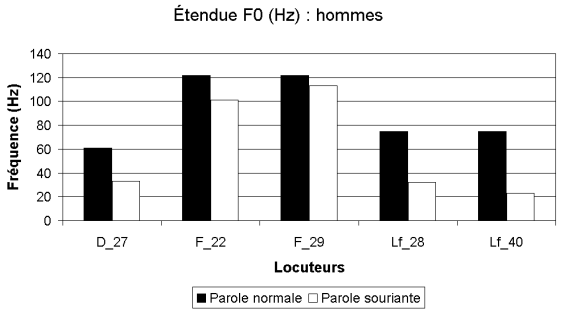


Figure 4 : Étendue de F0 (Hz) pour 5 énoncés produits par 3 locuteurs, condition neutre et sourire

Ces résultats nous rappellent ceux obtenus par Thibault [10] qui avait démontré que l'étendue de F0 est plus grande en situation de lecture qu'en parole spontanée. Une relation est possible avec nos résultats : l'étendue de F0 est plus grande en parole normale qu'en parole souriante. La lecture, chez Thibault [10], et la parole normale dans notre étude, correspondent aux conditions neutres (non marquées) ont une plus grande étendue; alors que la parole spontanée et la parole souriante, (conditions marquées), ont une plus petite étendue. Ceci suggèrerait un lien avec le style de parole, le style de discours et l'étendue de la F0. Ceci pourrait servir à un premier débroussaillage des contextes, des situations où apparaissent le rire et le sourire.

4.3. Limites de la recherche

La variation intra/inter sujets et certains impondérables nous empêche, pour l'instant, d'aboutir à des conclusions générales sur la parole souriante. Un des cas inattendus, pour tous les locuteurs, a été la propagation du sourire d'un énoncé *x* sur un énoncé *y* où les locuteurs, en lisant un énoncé neutre, continuaient de sourire en repensant à la caricature qu'ils venaient tout juste de voir. Conséquences : Il n'est pas possible d'analyser les mêmes énoncés pour tous les locuteurs et les énoncés des caricatures produits avec le sourire n'ont pas tous leur contrepartie neutre.

5. CONCLUSION

Nos données démontrent que le sourire est perceptible dans la parole. En plus du contenu linguistique lui-même, des informations paralinguistiques et extralinguistiques sont également transmises à travers les signaux de la parole (Traunmüller [11]). À ces signaux, il est impossible de soustraire les qualités individuelles et l'état d'un locuteur (sexe, âge, émotions, etc.). Lorsque le locuteur devient auditeur, il est intéressant de constater qu'il tient compte, inconsciemment, des qualités individuelles de son interlocuteur : « … in speech perception, listeners do not evaluate the acoustic cues directly but rather in a relational way, taking the personal properties of the speaker into account » (Traunmüller [11], p. 171) .

La moyenne et l'étendue de F0 offrent une piste intéressante si on les combine avec d'autres paramètres telles les situations de discours, les différents contextes. Nous avons remarqué que dans la parole souriante, il y a souvent des pauses, des hésitations, des faux départs, des reprises, etc. Ceci suggère que la facette pragmatique du sourire aurait un rôle non négligeable à jouer, ce qui pourrait nous aider éventuellement à bâtir une véritable typologie dans le domaine.

BIBLIOGRAPHIE

[1] V. Aubergé and M.-A. Cathiard. Can we hear the prosody of smile? *Speech Communication*, 40: 87-97, 2003.

[2] J.-A. Bachorowski,, M. J. Smoski and M. J. Owren. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 111 (3): 1582-1597, 2001.

[3] S. Chapleau. *L'année Chapleau 2003*. Boréal, Montréal, 2003.

[4] S. Chapleau. *L'année Chapleau 2004*. Boréal, Montréal, 2004.

[5] A. Hirson. Human Laughter – A Forensic Phonetic Perspective. In *Studies in Forensic Phonetics*, sous la dir. de A. Braun et J.-P. Köster, Wissenschaftlicher Verlag, Trier, 77-86, 1995.

[6] D. E. Mowrer, L. L. LaPointe and J. Case. Analysis of five acoustic correlates of laughter, *Journal of Nonverbal Behavior*, 11 (3): 191-199, 1987.

[7] R. Provine. *Le rire, sa vie, son œuvre*. Trad. de l'américain par J.-L. Fidel. Robert Laffont, Paris, 2003.

[8] M. Schröder, V. Aubergé and M.-A. Cathiard. Can we hear smile? In *Proc. Conf. on Spoken Language Processing*, volume 3, pages 559-562, 1998.

[9] V. C. Tartter. Happy talk: Perceptual and acoustic effects of smiling on speech, *Perception and Psychophysics*, 27 (1): 24-27, 1980.

[10] L. Thibault. Variations phonétiques et tonales en français québécois lu et spontané. Thèse de doctorat, Université du Québec à Montréal, 1998.

[11] H. Traunmüller. Conventional, Biological and Environmental Factors in Speech Communication: A Modulation Theory. *Phonetica*, 51: 170-183, 1994.

[12] J. Trouvain. Phonetic Aspects of "Speech-Laughs". In *Proc. of ORAGE, Orality and Gestuality Conference*, 634-639, 2001.

[13] J. Trouvain. Segmenting Phonetic Units in Laughter. In *Proc. of ICPhS*, 2793-2796, 2003.

Indices acoustiques de la coarticulation bidirectionnelle dans les séquences VCV en arabe

Mohamed Embarki

ICAR-Praxiling UMR 5191 CNRS-Montpellier III
Route de Mende, 34199 Montpellier Cedex 5, France
mohamed.embarki@univ-montp3.fr

Abstract

This study assessed anticipatory and carry-over coarticulation effects in contemporary standard Arabic. VCV pairs with non pharyngealized dental-alveolar consonants (/t/, /d/, /s/, /ð/ and their pharyngealized cognates (/tˤ/, /dˤ/, /sˤ/, /ðˤ/) were used. Each consonant was inserted in symmetric vocalic contexts [iCi], [uCu] and [aCa]). F2 was measured at V1mid, V1offset, V2onset and V2mid. The results showed carry-over effects with non pharyngealized consonants and anticipatory coarticulation in pharyngealized context.

1. Introduction

Le système de l'arabe standard contemporain (ASC) fonctionne avec 3 voyelles orales brèves (/i/, /u/, /a/) et trois longues (/iː/, /uː/, /aː/), lesquelles peuvent être décrites avec seulement quatre traits du modèle *SPE* ([+-haut], [+-arrière], [+-arrondi], [+-long]. Les articulations consonantiques engendrent une modification des voyelles, tant au niveau physiologique que physique (Al-Ani [1], Ghazali [2], Znagui [3]). Les consonnes pharyngalisées /tˤ/, /dˤ/, /sˤ/, /ðˤ/ sont décrites comme les plus influentes sur leur entourage vocalique. Réalisées avec une articulation principale dentale/alvéolaire et une articulation secondaire avec une rétractation du dos de la langue vers la paroi pharyngale (Al-Ani [1], Bonnot [4]), les consonnes pharyngalisées abaissent les voyelles contiguës fermées (/i/,/iː/, /u/, /uː/) et entraînent les voyelles ouvertes (/a/ et /aː/) vers une articulation postérieure, modifiant ainsi la fréquence de leurs deux premiers formants, F1 et F2 (Ghazali [2], Yeou [5]).

Comment cette variabilité est-elle représentée au niveau phonologique ? La littérature ne permet pas encore une réponse satisfaisante. Car dans le domaine arabe, les aspects cognitifs et physiques de la production de la parole continuent globalement d'être explorés de manière cloisonnée dans deux domaines qui n'interagissent que très rarement : d'un côté le domaine dont les objets premiers sont les unités discrètes et abstraites du système, les phonèmes ; et de l'autre celui confronté aux dimensions variables du signal de parole.

La coarticulation, un objet d'étude approchable des niveaux linguistique, moteur, acoustique, perceptif, peut apporter une meilleure connaissance non seulement de la production, mais aussi de la programmation et de la représentation de la parole en arabe.

2. Représenter la coarticulation

Farnetani et Recasens [6] montrent que les données de la littérature oscillent entre aspects universels et contraintes linguistiques spécifiques. D'une langue à l'autre, un même segment n'est aucunement influençable de manière similaire par les segments adjacents. Aussi, les segments de la même langue ne résistent-ils pas tous de la même manière à la coarticulation. A l'appui de multiples comparaisons à l'issue desquelles les consonnes [s] et [z] paraissaient les mieux résistantes aux sons contigus et non contigus, Bladon et Al-Bamerni [7] ont proposé le concept de résistance articulatoire (Articulatory Resistance), doublé d'un coefficient pour dépasser la représentation binaire des traits et pour mieux rendre compte de l'aspect gradué de la coarticulation. Le concept de sous-spécification (underspecification), autre aspect de représentation phonologique de la coarticulation, précise que les traits jouent plusieurs rôles aux niveaux lexical, phonologique et phonétique, et que certains traits superflus lexicalement et phonologiquement sont présents phonétiquement. Le modèle proposé par Keating [8] tient compte de la sous-spécification tant au niveau phonologique que phonétique. L'étude de Cohn [9] sur la nasalité montre que les voyelles de l'anglais qui sont sous-spécifiées par rapport à ce trait passent progressivement en fonction du contexte de [-nasal] à [+ nasal]. Clements [10] montre que trait tendu/relâché vient au renfort de l'opposition bref *vs* long dans certaines variétés du General American English (GAE), les voyelles longues réalisées tendues et les brèves relâchées. La quantité qui est phonologique est actualisée avec la tension, trait purement phonétique.

3. La coarticulation VCV

La coarticulation est décrite dans la littérature comme pouvant être anticipatoire (gauche-droite) reflétant une activité de pré programmation du segment, ou rémanente (droite-gauche) consécutive à des limitations mécaniques et à l'inertie des organes articulateurs (Daniloff et Hammarberg [11] ; Farnetani et Recasens [6] ; Recasens [12]). Plusieurs aspects ont été examinés, moteur, acoustique et perceptif. Du point de vue moteur, les éléments les plus étudiés concernent la coarticulation labiale, linguale, vélopharyngale et laryngale ; d'un point de vue acoustique, on y étudie la fréquence des formants et la durée. S'appuyant sur l'étude d'Öhman [13] à propos des séquences VCV montrant que la production des trois segments n'était pas une suite linéaire de gestes - les voyelles co-articulent l'une avec l'autre à l'instar des gestes pour une diphtongue sur lesquels les gestes

consonantiques sont surimposés- la littérature utilise le concept de coarticulation de voyelle à voyelle (V-to-V) pour désigner les effets de la voyelle sur les transitions trans-consonantiques. Pour ne présenter que la partie linguale, Recasens [11] montre que le degré de coarticulation de V-à-V dans les contextes VCV et VCVCV est dépendant de la surface de la langue mobilisée dans l'articulation de la consonne, plus le contact est important et moins il y a d'effets entre les voyelles. Selon le modèle de « degré de contraintes articulatoires » (DAC), Recasens et al. [14] indiquent que les consonnes qui requièrent une partie importante du dos de la langue sont plus contraintes que celles qui requièrent la pointe de la langue ou les lèvres, le degré de contrainte pour les occlusives va décroissant ([k]>[t]> [p]). Dans la séquence VCV où la consonne est moins contrainte, la coarticulation peut être bidirectionnelle. Toutefois, dans la séquence [ini] comparée à [ana], la voyelle [i] favorise la coarticulation rémanente de C-à-V, car avec une consonne non contrainte dentale/alvéolaire, [i] mobilisant le dos de la langue, bloque toute stratégie anticipatoire. Au niveau labial, les travaux sur le français ont montré que la labialité de V2 dans la séquence VCV est présente dès la fin de V1 (Sock et Vaxelaire. [15]), toutefois la coarticulation V-à-C et de C-à-V peut être aussi bien anticipatoire que rémanente, dépendant du trait la voyelle [+- arrondie] (Abry et Lallouache, [16]). Au niveau acoustique, les travaux de Sussman et ses collaborateurs ([17] et [18]) ont montré que les équations de locus sont indicatives du lieu d'articulation de la consonne. Modarresi et *al.* [19] ont exploré la bidirectionnalité de la coarticulation, anticipatoire et rémanente dans les séquences V.CV et VC.V. Les mesures de F2 prises aux frontières de C – F2 offset dans V1C et F2 onset dans CV2 – montrent que les effets de coarticulation rémanente dépassent les effets anticipatoires en syllabe fermée, alors que les effets anticipatoires ne dépassent les effets rémanents en syllabe ouverte qu'en contexte vocalique antérieur ([i] et [e]), avec le contexte vocalique postérieur ([u] et [ɔ]) les effets gauche-droite et droite-gauche se neutralisent.

Nous explorerons ici d'un point de vue acoustique la coarticulation V-à-C et C-à-V dans les séquences VCV en ASC et vérifieront l'impact de l'opposition consonantique pharyngalisé *vs* non pharyngalisé.

4. Méthodologie

Les séquences VCV qui sont analysées ici sont extraites d'un corpus de 24 mots en ASC. Nous n'avons pas pu faire d'opposition de contexte vocalique fixe *vs* changeant (Recasens et al. [14]), nous avons préféré un contexte symétrique, [iCi], [uCu], [aCa]. Le contexte consonantique choisi est dental/alvéolaire, car selon le concept de résistance articulatoire (Bladon et Al-Bamerni [7]), et outre le fait que ce contexte est celui qui jouit du degré de résistance le plus élevé parmi les consonnes, il est aussi celui qui exerce le plus d'effet coarticulatoire sur les segments adjacents (Farnetani et Recasens [20]). Le contexte consonantique choisi dans les séquences VCV est soit pharyngalisé ([tˁ], [dˁ], [sˁ] et [ðˁ]) ou non pharyngalisé ([t], [d], [s] et [ð]).

Huit locuteurs arabophones de sexe masculin, âgés de 25 à 40 ans, originaires de huit pays (Maroc, Algérie, Libye, Soudan, Liban, Jordanie, Arabie Saoudite et Koweït) ont été enregistrés. Les 24 mots ont été mis dans une phrase porteuse du type [qul...ljawm] (dis… aujourd'hui). Chaque locuteur a lu trois fois le corpus, un ensemble de 576 mots a été segmenté et étiqueté sous PRAAT [(8 consonnes x 3 voyelles x 3 répétitions x 8 locuteurs =576].

Les mesures de F2 ont été prises conformément à la littérature (Recasens [21], Modarresi et al. [19]) au milieu de V1 (V1mid) et à la fin de celle-ci (V1offset), au début et au milieu de V2 (respectivement V2onset et V2mid). Les mesures de V1mid et V2mid ont été prises à 50% du cycle temporel de la voyelle coïncidant souvent avec la partie stable de celle-ci ; les mesures de V1offset et V2onset ont été prises respectivement sur la dernière et la première résonances visible de F2.

4. Résultats

Les résultats portent sur 2304 mesures de F2, la moyenne a été calculée pour chacune des trois voyelles indépendamment du contexte consonantique, pharyngalisé ou non pharyngalisé (table 1 ci-après).

Table 1 : Valeurs moyennes de F2 (en Hz) chez 8 locuteurs arabophones (Moy=moyenne ; E-T=écart-type)

V		V1mid	V1offset	V2onset	V2mid
[i]	Moy.	2228	1857	1862	2115
	E-T	134	347	395	314
[u]	Moy.	927	1114	1127	969
	E-T	137	280	309	256
[a]	Moy.	1333	1330	1364	1363
	E-T	190	255	260	213

Des calculs ANOVA ont été effectués sur les points-clefs (V1mid *vs* V1offset ; V1offset *vs* V2onset ; V2onset *vs* V2mid ; V1mid *vs* V2mid) afin de vérifier l'égalité des variances indépendamment du point de mesure de F2 - partie stable (mid) ou onset/offset de la transition – et du contexte consonantique. Les effets sont très significatifs pour la voyelle [a], (a1mid *vs* a1offset [F (1, 191) =7.895 ; p<0.0001] ; a1offset *vs* a2onset [F (1, 191) =10.711 ; p<0.0001] ; a1mid *vs* a2mid [F (1, 191) =11.807 ; p<0.0001] ; a2onset *vs* a2mid [F (1, 191) =16.621 ; p<0.0001], autrement dit les effets sur la voyelle sont mineures. Les points-clefs de la voyelle [i] ne sont pas homogènes, deux sont hautement significatifs (i1offset *vs* i2onset [F (1, 191)=10.711 ; p<0.0001] et i2onset *vs* i2mid [F (1, 191)=5.162 ; p<0.0001]) et les deux autres non significatifs (i1mid *vs* i1offset [F (1, 191)=1.196 ; p=0.206] ; i1mid *vs* i2mid [F (1, 191) =1.181 ; p=0.220]. Hormis u1mid *vs* u1offset qui n'est pas significatif [F (1, 190)=1.102 ; p=0.329], les trois autres points-clefs sont très significatifs (u1offset *vs*

u2onset [F (1, 190)=8.476 ; p<0.0001] ; u1mid *vs* u2mid [F (1, 190)=1.980 ; p<0.001] ; u2onset *vs* u2mid [F (1, 190)=7.594 ; p<0.0001]). Les calculs ANOVA double-facteurs à mesures répétées avec les différents points-clefs comme variable dépendante et la consonne comme regresseur montrent des effets très significatifs. Sur 96 ANOVA (4 points-clefs x 3 voyelles x 8 consonnes), seulement 27 se sont révélées non significatives, celles portant sur un point de chaque voyelle (V1offset *vs* V2onset = 11 et V1mid *vs* V2mid = 9 contre V1mid *vs* V1offset = 5 et V2onset *vs* V2mid = 2). Il apert donc que le contexte consonantique exerce des effets coarticulatoires importants sur la fréquence de F2.

Modarresi et *al.* [19] ont évalué les effets de la coarticulation sur la base d'un Hz calculé entre V1offset et V2onset avec des voyelles fixes et changeantes. Notre corpus ne contient que des contextes vocaliques fixes (iCi, uCu, aCa). Aussi, le Hz sera calculé à partir de la somme des deux points-clefs de V1 et la somme des deux points-clefs de V2, i.e. Hz = [(V1mid + V1offset)-(V2onset+V2mid)]. Etant donné que l'onset et l'offset de la voyelle sont influençables par le contexte consonantique et la partie stable de la voyelle pas ou peu influençable, si Hz est positif, la coarticulation est anticipatoire (anticipatoire > rémanente) ; s'il est négatif, la coarticulation est rémanente (rémanente > anticipatoire) ; si Hz est compris entre -20 et +20, les deux effets se neutralisent (anticipatoire <> rémanente).

Les Hz ont été comparés en fonction de la voyelle et en fonction de la consonne. Les calculs ANOVA à double-facteurs à mesures répétées avec [V1mid +V1offset) - (V2onset+V2mid) comme variable dépendante et le facteur pharyngalisé *vs* non pharyngalisé comme regresseur montrent un effet hautement significatif pour les voyelles fermées [i] [F (1, 95)=0.924 ; p=0.03] et [u] [F (1, 95)=1.314 ; p=0.029] et faiblement significatif pour la voyelle ouverte [a] [F (1, 95)=0.811 ; p=0.04]. Les Hz moyens selon la nature de la voyelle montrent que la coarticulation rémanente est globalement majoritaire, elle n'est anticipatoire qu'avec [i] et [u] en contexte pharyngalisé.

Si l'on examine les données en tenant compte de la consonne et de la voyelle, la tendance est plus nuancée. En fonction de la nature de la consonne, pharyngalisée *vs* non pharyngalisée, les effets de coarticulation changent de direction (cf. table 2 ci-après). Les douze Hz moyens (3 voyelles x 4 consonnes) en contexte non pharyngalisé indiquent que les effets rémanents dépassent les effets anticipatoires, respectivement dans 9 cas contre 1, les deux cas restants sont neutres. En revanche, les effets anticipatoires priment les effets rémanents (9 contre 2) en contexte pharyngalisé. Les deux occlusives non pharyngalisées [t] et [d] s'accompagnent d'effets rémanents plus importants que les effets anticipatoires dans les trois contextes vocaliques [i], [u], [a]. Pour la consonne [s], les effets sont anticipatoires avec [i], rémanents avec [u] et neutralisés avec [a]. Pour la consonne [ð], la coarticulation est rémanente avec [i] et [a] et neutre avec [u]. Ces résultats sont conformes à la littérature, Modarresi et *al.* [19] ayant montré que les consonnes alvéolaires en syllabe ouverte montraient davantage d'effets rémanents que les labiales ou les vélaires. Recasens [14] a montré que la voyelle [i] comparée à [a], dans la séquence VCV favorise la coarticulation rémanente de C-à-V puisque la consonne dentale/alvéolaire est non contrainte et que [i] requiert l'élévation du *dorsum*, bloquant toute stratégie anticipatoire. Nos résultats montrent que [i] est associé dans 3 contextes sur 4 à des effets rémanents ; le cas de [s] est conforme au principe de résistance articulatoire (Bladon et Al-Bamerni [7]). Le cas de [isi] où les effets anticipatoires priment les rémanents n'est pas étonnant, car selon Farnetani et Recasens [20] le segment à résistance articulatoire élevée exerce aussi une influence élevée sur les segments contigus.

Table 2 : Hz = [(V1mid + V1offset)-(V2onset+V2mid)] et sens de coarticulation dans les séquences VCV ([An>Re]= anticipatoire>rémanente ; [Re>An]= rémanente>anticipatoire ; [An<>Re]=équilibré) (*=p<0.01 ; **=p<0.001 ; #=non significative)

C	V	Hz	An >Re	Re >An	An <> Re	C	V	Hz	An >Re	Re >An	An <> Re
t	i	-282**	4	**20**	1	tˤ	i	199**	**14**	8	0
	u	-233**	4	**20**	0		u	159**	**16**	3	2
	a	-119**	5	**19**	0		a	0**	9	10	**5**
			0	**3**	0				**2**	0	**1**
d	i	-145**	3	**17**	4	dˤ	i	330**	**18**	6	0
	u	-494**	1	**22**	1		u	107**	**18**	6	0
	a	-186*	1	**22**	0		a	-91*	4	**20**	0
			0	**3**	0				**2**	**1**	0
s	i	171#	**14**	10	0	sˤ	i	387#	**16**	9	0
	u	-112**	6	**15**	3		u	101**	**15**	8	**1**
	a	-16#	13	11	**0**		a	42#	**14**	8	2
			1	**1**	**1**				**3**	0	0
ð	i	-140**	2	**22**	2	ðˤ	i	357**	**17**	6	0
	u	-2#	15	9	**1**		u	50#	**15**	9	0
	a	-69#	4	**20**	0		a	-73#	7	**17**	0
			0	**2**	**1**				**2**	**1**	0

Pour les consonnes pharyngalisées, les effets sont anticipatoires dans les trois contextes vocaliques pour [sˤ], et seulement avec [i] et [u] pour les consonnes [tˤ], [dˤ], [ðˤ]. La voyelle [a] semble bloquer les effets anticipatoires de la consonne pharyngalisée. Cela est dû à une plus grande compatibilité entre l'aperture de la voyelle et la rétractation du dos de la langue vers la paroi pharyngale (Al-Ani [1], Ghazali [2], Znagui [3], Bonnot [4]).

La fréquence moyenne des quatre points-clefs des voyelles (V1mid ; V1offset ; V2onset ; V2mid) en fonction du contexte consonantique non pharyngalisé et pharyngalisé révèle que quand la consonne est non pharyngalisée les effets de coarticulation anticipatoires et rémanents s'arrêtent avec la transition de V2. En revanche, quand la consonne est pharyngalisée, ses effets anticipatoires sont très marqués aux frontières de la consonne (V1offset et V2onset) et se prolongent jusqu'au

milieu de V2. Il est fort probable que les mêmes effets puissent apparaître bien avant V1 et se prolongent au-delà de V2.

Conclusion

Nous avons présenté dans cette étude exploratoire quelques aspects acoustiques partiels de la coarticulation VCV en contexte consonantique non pharyngalisé *vs* pharyngalisé. Les résultats montrent principalement les effets anticipatoires de la consonne pharyngalisée et les effets rémanents des consonnes dentales/alvéolaires non pharyngalisés. Ces indices acoustiques témoignent de stratégies différentes de programmation et de production de cette opposition consonantique. Bien que non traités ici, ces aspects révèlent indirectement une représentation linguistique différente de cette opposition. Les voyelles de l'arabe sous-spécifiées par rapport au trait [+- bas], passent en contexte pharyngalisé progressivement de [+ haut] à [-haut], de [0] à [-bas] pour [i] et [u] ; de [-arrière] à [+arrière] pour la voyelle [a], celle-ci nécessitant les traits [-haut] et [+bas] pour éviter la confusion avec [u].

Bibliographie

[1] S.H. Al-Ani. *Arabic phonology*. Mouton, The Hague, 1970.

[2] S. Ghazeli. *Back consonants and backing articulation in Arabic*. Ph.D. Dissertation, University of Texas, 1977.

[3] I. Znagui. *Etudes phonétique et perceptive des voyelles de l'arabe standard moderne*. Thèse de Doctorat, université Paris III, 1995.

[4] J.-F. Bonnot. *Contribution à l'Etude des Consonnes Emphatiques de l'Arabe à partir de Méthodes Expérimentales*. Thèse de Doctorat de 3ème cycle, Université des Sciences Humaines de Strasbourg, 1976.

[5] M. Yeou. Locus equations and the degree of coarticulation of Arabic consonants. *Phonetica*: 54, 187-202.

[6] E. Farnetani and D. Recasens. Coarticulation models in recent speech production theories. In: W.J. Hardcastle and N. Hewlett (eds.), *Coarticulation. Theory, data and techniques*. Cambridge University Press, Cambridge, UK, pages 31-65, 1999.

[7] R.A.W. Bladon and A. Al-Bamerni. Coarticulation resistance in English /l/. *Journal of Phonetics*: 4, 137-150, 1976.

[8] A.C. Cohn. Phonetic and phonological rules of nasalization. *UCLA Working Papers in Phonetics*, 76, 1990.

[9] P.A. Keating. Universal phonetics and the organization of grammars. In: V. Fromkin (ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged*. Academic Press, Orlando, pages 115-132, 1985.

[10] G.N. Clements. Les diphtongues brèves en anglais : fonction phonétique du trait tendu/relâché. In: J.-P. Angoujard et S. Wauquier-Gravelines (eds.), *Phonologie. Champs et perspectives*, ENS Editions, Lyon, pages 35-55, 2003.

[11] R. Daniloff and R. Hammarberg. On defining coarticulation. *Journal of Phonetics*: 1, 239-248, 1973.

[12] D. Recasens. Vowel-to-vowel coarticulation in Catalan VCV sequences. *JASA*: 73, 1624-1635, 1984.

[13] S. Öhman. Coarticulation in VCV utterances: spectrographic measurements. *JASA*: 39, 151-168, 1966.

[14] D. Recasens, M.D. Pallarès and J. Fontdevilla. A model of lingual coarticulation based on articulatory constraints. *JASA*: 102, 544-561, 1997.

[15] R. Sock et B. Vaxelaire (eds.). *L'anticipation à l'horizon du présent*. Mardaga, Sprimont, 2004.

[16] C. Abry et M.T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français. In : *Bulletin du Laboratoire de la Communication Parlée*, volume 3, pages 85-99, 1995.

[17] H.M. Sussman, H.A. McCaffrey and S.A. Mathews. An investigation of locus equations as a source of relational invariance for stop place of articulation. *JASA*: 90, 1309-1325, 1991.

[18] H.M. Sussman, K. Hoemeke and F. Ahmed. A cross-linguistic investigation of locus equations as a relationally invariant descriptor of place of articulation. *JASA:* 94, 1256-1268, 1993.

[19] G. Modarresi, H.M. Sussman, B. Lindblom and E. Burlingame. An acoustic analysis of the bidirectionality of coarticulation in VCV utterances. *Journal of Phonetics*: 32, 291-312, 2004.

[20] E. Farnetani and D. Recasens. Anticipatory consonant-to-vowel coarticulation the production of VCV sequences in Italian. *Language and Speech*: 36, 279-302, 1993.

[21] D. Recasens. Acoustic analysis. In: W.J. Hardcastle and N. Hewlett (eds.), *Coarticulation. Theory, data and techniques*. Cambridge University Press, Cambridge, UK, pages 322-336, 1999.

Equation de locus comme indice de distinction consonantique pharyngalisé vs non pharyngalisé en arabe

Mohamed Embarki, Christian Guilleminot** & Mohamed Yeou****

*ICAR-Praxiling UMR 5191 CNRS-Montpellier III
Route de Mende, 34199 Montpellier Cedex 5, France
mohamed.embarki@univ-montp3.fr
**Centre Tesnières (EA 2283), Université de Franche-Comté, Besançon
christian.guilleminot@univ-fcomte.fr
*** Université Chouaib Doukkali, El-Jadida, Maroc
m_yeou@yahoo.com

ABSTRACT

Locus equations are linear regression functions derived by relating F2 onsets of different vowels to their corresponding steady states. This paper purports to investigate if locus equations can be strong phonetic descriptors of the consonantal contrast between pharyngealized and non-pharyngealized consonants in Arabic. Eight male Arabic speakers from eight different Arabic countries produced 24 #CV# tokens, where C was either non-pharyngealized [t], [d], [s] and [ð] or pharyngealized [ð] [tˁ], [dˁ], [sˁ] and [ðˁ]. Each consonant was followed by one of the three vowels [i], [u] and [a].

1. INTRODUCTION

La diversité des parlers populaires arabes utilisés dans l'aire arabophone retient l'attention des chercheurs depuis longtemps. Sur le plan phonétique, la variabilité sensori-motrice a débouché sur un classement en zones géographiques plus ou moins homogènes (Barkat [1] ; Sabhi [2]). Les traits phonologiques des parlers maternels sont détectés dans la production en arabe standard contemporain (ASC) et deviennent même des indices de reconnaissance régionale, tant au niveau acoustique (Sabhi [2]), qu'au niveau perceptif (Barkat [1]). Néanmoins, ces parlers possèdent des traits phonologiques qui font incontestablement leur unité. Le trait d'opposition consonantique pharyngalisé *vs* non pharyngalisé est à ce titre unificateur des parlers arabes, et plus largement de tout le groupe sémitique (Catineau [3]). Or, l'actualisation en ASC de cette opposition consonantique n'en demeure pas moins étroitement liée à l'origine géographique du locuteur (Sabhi [2]).

2. COARTICULATION ET ÉQUATION DE LOCUS

L'équation de locus, une régression linéaire obtenue à partir de la relation entre la fréquence de F2 au début de la voyelle (F2onset) sur l'axe des ordonnées et la fréquence de F2 à la partie stable (F2mid) sur l'axe des abscisses - F2onset =k*F2mid+c (où *k* et *c* sont la pente et l'ordonnée de la fonction de l'intersection-y) - a été utilisée à l'origine par Lindblom [4] comme indicatrice du degré de coarticulation entre la consonne et la voyelle. La pente de l'équation de locus variant entre les extrema 0-1 témoigne de la coarticulation entre ces deux segments au sein de la syllabe. Une pente relativement plate est indicatrice d'un minimum de coarticulation entre les deux segments, F2onset étant dans ce cas insensible à la nature de la voyelle qui suit (résistance coarticulatoire maximale de l'articulation de la consonne aux effets de la voyelle) ; une pente relativement forte est indicatrice d'un maximum de coarticulation entre les deux segments, F2onset et F2mid ont la même fréquence (résistance coarticulatoire minimale de l'articulation de la consonne). Sussman et collaborateurs [5 et 6] ont confirmé que les équations de locus sont un indice important de lieu d'articulation car leurs pentes varient en fonction de ce dernier: /g/ > /b/ > /d/. Les auteurs ont trouvé que la consonne vélaire a une pente à peine plus forte que celle de la consonne labiale, l'intersection-y est plus faible pour cette dernière ; la consonne dentale présente une valeur d'intersection-y élevée mais une pente plus plate. D'autres chercheurs (Krull [7 et 8] ; Fowler [9]) ont montré que les pentes de ces équations indiquent plutôt le degré de coarticulation.

La validité du concept de l'équation du locus a été confirmée dans plusieurs langues comme le thaï, l'urdu, et l'arabe égyptien (Sussman et al. [6]), le français, l'anglais américain et le suédois (Molis et coll. [10]), l'anglais américain et le persan (Modarresi et al. [11]). Tabain et Butcher [12] ont expérimenté l'équation de locus dans la comparaison de deux langues aborigènes d'Australie très proches, le yanyuwa, le yindjibarndi, avec l'anglais australien. Les résultats de la littérature souvent similaires penchent pour une distinction nette entre deux groupes : un groupe de consonnes dentales et alvéolaires et un groupe de consonnes labiales et vélaires.

3. LES CONSONNES PHARYNGALISÉES

Les consonnes pharyngalisées en ASC se distinguent de leurs correspondantes non pharyngalisées sur les plans moteur, acoustique et perceptif (Bonnot [13] ; Yeou [14]). Sussman et al. [6] ont montré que les consonnes pharyngalisées se distinguent de leurs

correspondantes par une pente plus faible et par des valeurs d'intersection-y plus basses. Les résultats de Yeou [14] sur l'ASC ont montré que les équations de locus permettent de distinguer les consonnes non pharyngalisées [t], [d], [s] et [ð] des consonnes pharyngalisées [ð] [tˁ], [dˁ], [sˁ] et [ðˁ], lesquelles émergent comme une classe distincte ayant les pentes les plus faibles et résistant à la coarticulation des voyelles adjacentes. Etant donné que l'ASC n'est pas une langue maternelle et considérant que l'équation de locus apparaît dès l'émergence de la coarticulation avec l'apparition des premiers mots, vers un an (Sussman et al. [15]), il est possible de détecter des différences de coarticulation en intégrant des locuteurs issus de régions différentes du Monde arabe. Notre hypothèse est que des locuteurs arabophones originaires de pays différents présenteront des équations de locus différentes, aussi bien entre groupes de consonnes (pharyngalisées vs non pharyngalisées) qu'entre consonnes du même groupe.

4. MÉTHODOLOGIE

Ont participé à cette expérience huit locuteurs arabophones originaires de huit pays arabes différents (Maroc, Algérie, Libye, Soudan, Liban, Jordanie, Arabie Saoudite et Koweït). Tous les locuteurs sont de sexe masculin, âgés de 25 à 40 ans et suivant des études à l'université de Franche-Comté (Besançon) et de Paul-Valéry (Montpellier). Chaque locuteur a lu trois fois le corpus de 24 mots en ASC où la syllabe CV qui fait l'objet de cette étude est médiane, [#CV#]. La consonne [C] est occupée en contexte non pharyngalisé par [t], [d], [s] ou [ð] et en contexte pharyngalisé par [tˁ], [dˁ], [sˁ] ou [ðˁ] ; la voyelle [V] est occupée par [i], [u] ou [a]. Les mots du corpus ont été insérés dans une phrase porteuse du type [qul...ljawm] (dis… aujourd'hui). Un ensemble de 576 mots a été segmenté et étiqueté sous PRAAT [(8 consonnes x 3 voyelles x 3 répétitions x 8 locuteurs =576]. Tous les locuteurs ont été enregistrés selon la même procédure, au laboratoire de phonétique de Besançon et à l'Atelier des Sciences du Langage de Montpellier. Les lexèmes ont tous été segmentés et les mesures de F2onset et F2voyelle ont été calculées manuellement. Les valeurs de F2onset ont été relevées sur la valeur la plus ample du début de la voyelle ; les valeurs de F2voyelle ont été relevées au milieu de la voyelle et dans la mesure du possible sur une partie stable.

5. RÉSULTATS

Les résultats du groupe consignés dans table 1 (ci-dessous) sont relativement conformes à la littérature, la pente des consonnes pharyngalisées est plus faible comparativement à celles des consonnes non pharyngalisées, la valeur de l'intersection-y a tendance à y être plus faible.

5.1. Résultats globaux

Table 1 : valeur de l'intersection-y (inter-y), de la pente et du coefficient de régression pour 8 locuteurs.

	Non pharyngalisé				pharyngalisé			
C	t	d	s	ð	tˁ	dˁ	sˁ	ðˁ
Inter-y	53 0.67	57 8.67	52 4.33	41 0.50	57 0.33	47 8.83	32 5.14	439 .17
pente	0.750	0.662	0.752	0.741	0.473	0.540	0.649	0.487
R^2	0.949	0.884	0.848	0.903	0.883	0.868	0.885	0.839

Les calculs ANOVA à double-facteurs à mesures répétées avec F2onset comme variable dépendante et F2voyelle comme régresseur montrent un effet significatif de la pharyngalisation sur la pente [F(1, 64) =27.03 ; p<0.001]. La valeur moyenne de la pente est plus forte dans [t] (0.75) que dans [tˁ] (0.47), les différences sont significatives [F(1, 7)=21.23 ; p=0.002] (cf. figures n° 1 et 2 ci-après). Entre [ð] et [ðˁ] des différences nettes de pente sont observables, respectivement 0.74 contre 0.48, les différences sont significatives [F(1, 7)=67.03, p<0.001]. Si la consonne voisée [d] présente une pente également plus forte que sa correspondante pharyngalisée [dˁ], respectivement 0.66 et 0.54, les différences ne sont pas significatives [F(1, 7)=4.15, p=0.081] (cf. figures n° 3 et 4 ci-après). Il en est de même pour les fricatives alvéolaires [s] et [sˁ], la pente de la consonne pharyngalisée est relativement plus plate (0.64 contre 0.75 pour [s]), les différences ne sont pas significatives non plus [F(1, 7) =5.86, p=0.52].

Si globalement, les moyennes de l'intersection-y des consonnes pharyngalisées sont plus basses que celles des consonnes non pharyngalisées, l'ANOVA n'a pas montré d'effet significatif de la pharyngalisation [F(1, 64)=0.39 ;p=0.54]. Comme le montre la figure 5, deux groupes homogènes de consonnes émergent quand on choisit de présenter en nuages de points les valeurs de l'intersection-y et les valeurs de pente : les consonnes non pharyngalisées ont des valeurs de pente globalement plus élevées que leurs correspondantes pharyngalisées, les premières ont une densité de valeurs importante entre 0.7 et 0.9, les secondes entre 0.4 et 0.6. Si le graphique montre qu'il n'existe pas de réelle zone de chevauchement entre les deux groupes de consonne, [sˁ] présente cependant des valeurs qui s'intègrent parfaitement dans le groupe des consonnes non pharyngalisées.

Pour les deux groupes de consonnes, Yeou [16] a trouvé des valeurs de pente différentes dans la production en ASC de 9 locuteurs originaires du

Maroc ([tˤ]=0.37, [dˤ]=0.31, [sˤ]=0.35 et [ðˤ]=0.22 ; [t]=0.66, [d]=0.48, [s]=0.56 et [ð]=0.46). Outre les différences statistiquement significatives entre les consonnes non pharyngalisées et leurs correspondantes pharyngalisées, Yeou [16] a trouvé des différences significatives entre [t], [d] et [s]. Si les équations de locus des consonnes non pharyngalisées dans notre étude sont légèrement différentes ([t]=0.75, [d]=0 .66, [s]=0.75 et [ð]=0.74), les différences ne sont cependant pas significatives. Dans le groupe des consonnes pharyngalisées, la comparaison montre que les pentes moyennes de [sˤ] (0.64) se distinguent de celles des autres consonnes pharyngalisées ([tˤ]=0.47, [dˤ]=0.54, et [ðˤ]=0.48), les différences étant statistiquement significatives. Outre l'absence de différences significatives entre d'une part certaines consonnes non pharyngalisées et leurs correspondantes pharyngalisées et d'autre part entre certaines consonnes du même groupe, nos résultats révèlent des équations de locus élevées pour des consonnes alvéolaires. Cette différence est en partie liée à l'endroit où sont prises les valeurs de F2onset, première résonance de F2 et non dans le burst.

5.2. Variabilité inter-locuteur

Les résultats globaux cachent une extrême variabilité inter et intra locuteur. Les calculs ANOVA à mesures répétées montrent un effet significatif de la pente selon le type de consonnes [F(3, 39)=2.10 ; p=0.12], ce qui révèle une très grande variabilité intra-sujet et cette variabilité est très significative. La variabilité inter-locuteur est, elle aussi, importante. Elle est sans aucun doute accentuée par l'origine géographique des locuteurs (figures 6 et 7). Toutefois, en l'absence de nombre suffisant de locuteurs par région, les données présentées dans notre ne peuvent être qu'indicatives.

CONCLUSION

Comment les Arabophones traitent la variabilité de contexte phonétique et retrouvent dans le signal de parole des représentations stables, telle est la question qui est posée *in fine* dans cette étude utilisant l'équation de locus comme degré de coarticulation au sein de la syllabe entre la consonne (pharyngalisée vs non pharyngalisée) et la voyelle adjacente. Nous avons montré que cette équation permet de distinguer les deux groupes de consonnes. Si nous n'avons observé des différences statistiquement significatives que sur deux couples de consonnes sur quatre, les consonnes pharyngalisées présentent des valeurs de pente relativement basses. L'équation de locus ne permet pas des distinctions nettes et ordonnées à l'intérieur de chaque groupe de consonnes. Globalement, notre étude révèle des équations de locus élevées pour des consonnes dentales/alvéolaires, pharyngalisées ou non pharyngalisées. Là où la littérature indique des valeurs autour de 0.50 pour le groupe de consonnes non pharyngalisées, nos résultats sont autour d'une moyenne de 0.75. Et là où l'étude de Yeou [14] révèle une moyenne de pente pour les consonnes pharyngalisées autour de 0.3, nos résultats sont autour de 0.57, avec une pointe à 0.64 pour [sˤ]. Une des raisons tient à la méthode de mesure : la résonance de F2 dans l'explosion ou dans la friction. Cette méthode a été utilisée dans Modarresi et al. [15]. Elle génère des pentes moins raides pour [t] et [d], respectivement 0.23 et 0.24. Son application reste difficile dans le cas de [s], ce qui explique les valeurs élevées de cette consonne dans Yeou [14]. Une autre raison tient à l'utilisation de voyelles brèves qui sont connues pour donner des pentes plus élevées.

La variabilité inter-sujet de l'équation de locus esquisse deux tendances qui restent à confirmer. La première concerne la relation variable au sein de la syllabe entre C et V en fonction du dialecte maternel. La seconde concerne le groupe de consonnes dentales/alvéolaires qui est en train d'évoluer vers une résistance moindre des consonnes aux effets de la voyelle, montrant ainsi un chevauchement plus grand des gestes articulatoires de la consonne et de la voyelle.

BIBLIOGRAPHIE

[1] M. Barkat. Détermination d'indices acoustiques robustes pour l'identification automatiques des parlers arabes. *Langues et Linguistique* : 7, 47-75, 2001.

[2] N. Sabhi. La variabilité dialectale arabe peut-elle être un moyen de reconnaissance de l'origine géographique ? Les fricatives interdentales, outils d'identification. *Revue Parole, 2*, 161-181, 1997.

[3] J. Cantineau. *Etude de Linguistique Arabe*, Klincksieck, Paris, 1960.

[4] B. Lindblom. On vowel reduction. *Report 29, The Royal Institute of Technology, Speech Transmission Laboratory*, Stockholm, 1963.

[5] H.M. Sussman, H.A. McCaffrey and S.A. Mathews. An investigation of locus equations as a source of relational invariance for stop place of articulation. *JASA*: 90, 1309-1325, 1991.

[6] H.M. Sussman, K. Hoemeke and F. Ahmed. A cross-linguistic investigation of locus equations as a relationally invariant descriptor of place of articulation. *JASA:* 94, 1256-1268, 1993.

[7] D. Krull. Acoustic properties as predictors of perceptual responses: a study of Swedish voiced stops. *Perilus:* 7, 66-70, 1988.

[8] D. Krull. Second formant locus patterns and consonant-vowel coarticulation in sponta-neous speech. *Perilus:* 10, 87-108, 1989.

[9] CA. Fowler. Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics:* 55, 597-610, 1994.

[10] M.R. Molis, B. Lindblom, W. Castelman and R. Carré. Cross-language analysis of VCV coarticulation. *JASA*: 95, 2925, 1994.

[11] G. Modarresi, H.M. Sussman, B. Lindblom and E. Burlingame. Locus equation encoding of stop place: revisiting the voicing/VOT issue. *Journal of Phonetics*: 33, 101-113, 2005.

[12] M. Tabain and A. Butcher. Slope values as acoustic measures of coarticulation: a cross-language comparison of stop consonants. Journal of Phonetics: 27, 333-357, 1999.

[13] J.-F. Bonnot. *Contribution à l'Etude des Consonnes Emphatiques de l'Arabe à partir de Méthodes Expérimentales.* Thèse de Doctorat de 3ème cycle, Université des Sciences Humaines de Strasbourg, 1976.

[14] M. Yeou. Locus equations and the degree of coarticulation of Arabic consonants. *Phonetica*: 54, 187-202.

[15] H/M. Sussman, K. Hoemeke and H. McCaffrey. Locus equations as an index of coarticulation and place of articulation distinctions in children. *Journal of Speech and Hearing Research:* 35, 397-420, 1992.

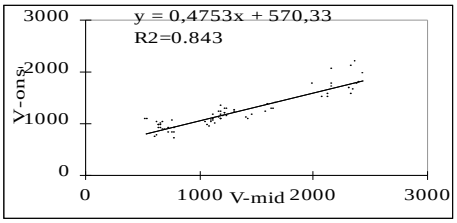


Figure 1 : Ligne de régression et valeurs de F2onset et F2voyelle de la consonne [tˁ]

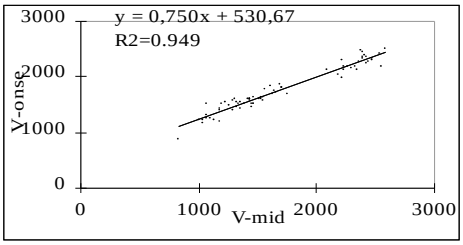


Figure 2 : Ligne de régression et valeurs de F2onset et F2voyelle de la consonne [t]

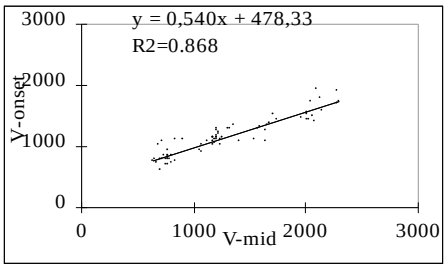


Figure 3 : Ligne de régression et valeurs de F2onset et F2voyelle de la consonne [dˁ]

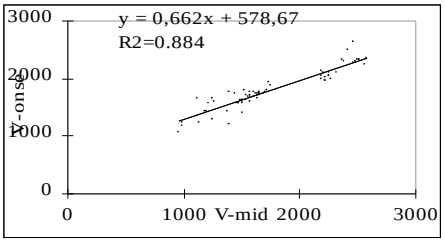


Figure 4 : Ligne de régression et valeurs de F2onset et F2voyelle de la consonne [d]

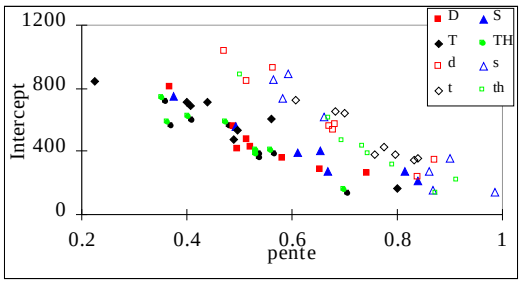


Figure 5 : valeurs de l'intersection-y (ordonnées) et valeur de la pente (abscisses) des 8 consonnes (D=[dˁ], T=[tˁ], S=[sˁ], TH=[ðˁ] et th=[ð]).

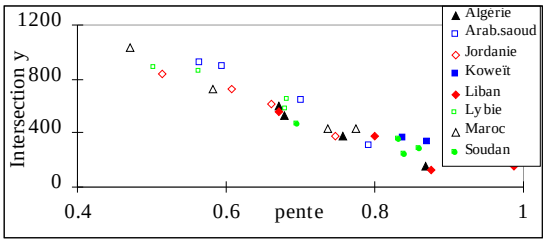


Figure 6 : intersection-y et pente pour les 4 consonnes non pharyngalisées chez les 8 locuteurs.

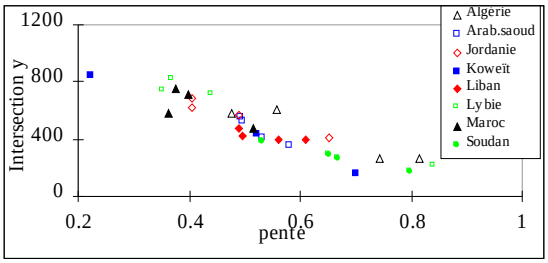


Figure 7 : l'intersection-y (ordonnées) et la valeur de la pente (abscisses) pour les 4 consonnes pharyngalisées chez les 8 locuteurs.

Paramétrisation de la Parole basée sur une Modélisation des Filtres Cochléaires: Application au RAP

Zied Hajaiej, Kais Ouni, Noureddine Ellouze

Laboratoire des Systèmes et Traitement du Signal (LSTS)
Ecole Nationale d'Ingénieurs de Tunis, BP 37, Le Belvédère, 1002 Tunis, Tunisie
Hajaiej_zied@yahoo.fr
(Kais.ouni, N. Ellouze@enit.rnu.tn)

ABSTRACT

Signal processing front end for extracting the feature set is an important stage in any speech recognition system. The optimum feature set is still not yet decided. There are many types of features, which are derived differently and have good impact on the recognition rate. This paper presents one more successful technique to extract the feature set from a speech signal, which can be used in speech recognition systems. Our technique based on the human auditory system characteristics and relies on the gammachirp filterbank to emulate asymmetric frequency response and level dependent frequency response. For evaluation a comparative study was operated with standard MFCC and PLP.

1. INTRODUCTION

Dans les deux dernières décennies, les modèles numériques appliqués au système auditif périphérique ont gagné une popularité croissante en traitement des signaux de parole, en particulier en reconnaissance de la parole. Par ailleurs, les filtres roex " rounded exponential " présentent une bonne approximation des données expérimentales auditives [4], sous les hypothèses simplificatrices que les filtres auditifs sont symétriques sur une échelle logarithmique, et que leur étalement loin de la fréquence centrale fc est négligé. Néanmoins, ils sont définit dans le domaine spectral ce qui rend difficile leur implémentation selon le schéma conventionnel des structures de bancs de filtres auditifs. Pour palier cet inconvénient, un modèle temporel a été proposé pour la première fois par Johannesma [1], appelée gammatone. Ce modèle a l'avantage d'être définie par une réponse impulsionnelle temporelle, c'est un filtre à bande critique obéissent à la mesure de Bark et se présente sous la forme d'une enveloppe de type gamma modulée. Le filtre gammatone présente une enveloppe spectrale symetrique, or les données psychoacoustiques optent pour une enveloppe non symétrique où le degré d'asymétrie dépend du niveau sonore [5], Irino et Patterson ont proposé un nouveau modèle du filtre auditif qui dérive de la fonction gammatone, appelé gammachirp, pour introduire une dépendance vis à vis du niveau d'intensité du stimulus sonore appliqué.

Cette dépendance se présente sous la forme d'un paramètre supplémentaire dans l'expression de la gammatone qui génère l'asymétrie du spectre d'amplitude. Dans ce papier nous proposons en premier lieu deux techniques de paramétrisation des signaux de parole basées sur un banc de filtres gammachirp qui imite le comportement spectral de la cochlée, en suivant la démarche utilisée dans les techniques MFCC et PLP. En second lieu, l'approche adoptée pour l'étude de la validité des deux techniques proposées ainsi que leur évaluation par rapport aux techniques de paramétrisation standards MFCC et PLP. Le système de reconnaissance adopté pour cette étude est celui de HTK basé sur les modèles de Markov cachés HMM.

2. LES TECHNIQUES DE PARAMETRISATION

Il existe dans la littérature une grande variété de technique de parametrisation des signaux de la parole, nous citons les plus important qui on révolutionne en quelque sorte le domaine de la reconnaissance de parole a savoir MFCC et PLP.

2.1. Coefficients mel-cepstre (MFCC)

Cette technique consiste à calculer les coefficients cepstraux sur une échelle en Mel qui se rapproche de la perception fréquentielle de l'oreille. Après l'application d'une transformée de Fourier à court terme, l'énergie est calculée dans des brandes critiques modélisées par des filtres triangulaires quant à l'échelle des amplitudes est exprimée en décibels. L'échelle des fréquences quant à lui est exprimée en Mel. Le cepstre est ensuite calculé par l'expression suivante :

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{N} \left(\log S_k \cos\left[n \left(k - \frac{1}{2} \right) \frac{\pi}{k} \right] \right) \qquad (1)$$

Avec k=1,…,N et S_k représentant l'énergie correspondante après filtrage par un $k^{ième}$ filtre triangulaire.

2.2.L'analyse prédictive linéaire perceptuelle (PLP)

La paramétrisation par prédiction linéaire (LPC) à pour principal défaut le fait d'estimer uniformément le spectre sur toutes les fréquences de la bande audible. Ainsi il est possible que certains détails spectraux ne soient pas pris en compte par la technique LPC ou encore qu'ils prennent une importance majeure sans qu'ils soient physiologiquement pris en compte par l'oreille. En effet, la technique PLP [9,11] permet de résoudre ce problème : l'analyse opérée par cette technique a pour but d'estimer des paramètres d'un filtre autorégressif tout pôle, modélisant au mieux le spectre auditif.

3. FILTRE GAMMACHIRP

Le filtre gammachirp est utilisé dans la recherche psychoacoustique comme étant un modèle fiable du filtre cochléaire. Il est défini dans le domaine temporel par la partie réelle de la fonction $g_c(t)$ [1, 3, 5].

$$g_c(t) = at^{n-1}\exp(-2\pi bERB(f_r)t)\times\exp(j2\pi f_r t + jc\ln t + jc\varphi) \quad (2)$$

Avec t>0, a paramètre de normalisation d'amplitude, f_r la fréquence de modulation, n l'ordre du filtre, $bERB(f_r)$ un paramètre définissant l'enveloppe du filtre. L'ERB représente quant à lui la largeur de bande rectangulaire équivalente [3,6].

$$ERB(f_r) = 24.7 + 0.108 f_r \quad (3)$$

c représente un facteur introduisant l'asymétrie de ce filtre et φ la phase initiale, ln t est un logarithme népérien de temps, La figure 1 donne un exemple de réponse impulsionnelle du filtre gammachirp.

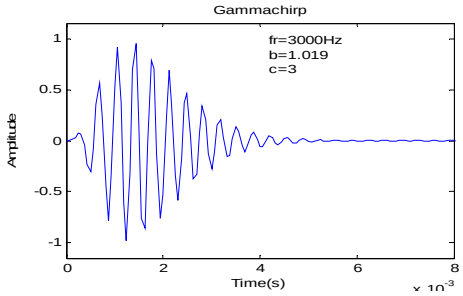


Figure 1 : Exemple de réponse impulsionnelle du filtre gammachirp.

La transformée de Fourier de la réponse impulsionnelle est donnée par l'équation suivante [3] :

$$G_c(f) = \frac{a|\Gamma(n+jc)|}{\Gamma(n)} \cdot \frac{\Gamma(n)}{|2\pi\sqrt{(b.ERB(f_r))^2 + (f-f_r)^2}|^n} e^{c\theta} \quad (4)$$

$$|G_c(f)| = a_\Gamma |G_T(f)| . e^{c\theta(f)} \quad (5)$$

$$\theta(f) = \arctan\left(\frac{f - f_r}{bERB(f_r)}\right) \quad (6)$$

Le transfert de gammachirp se présente comme le produit du transfert de la gammatone $G_T(f)$ par une fonction de transfert appelée fonction d'asymétrie $e^{c\theta(f)}$. Le degré d'asymétrie dépend de c, si c est négatif la fonction de transfert $e^{c\theta(f)}$ se comporte comme un filtre passe-bas et dans le cas ou C est positif elle se comporte comme un filtre passe-haut. Les études psychoacoustiques montre que c est fortement dépendant de la puissance du signal. En effet le paramètre c est relié à la puissance du signal par une expression de type $c=3.38-0.107P_s$ [3], avec P_s puissance de signal d'entré. Le pic de fréquence f_p de ce spectre est défini pour $G'(f_p)=0$. Ce pic est décalé de f_r par [4, 5].

$$f_p = \frac{f_r + cbERB(f_r)}{n} \quad (7)$$

Ce décalage est dû au paramètre c introduit dans l'expression de la réponse impulsionnelle gammachirp g_c. Ce décalage ainsi que l'asymétrie du spectre de la gammachirp représente une approximation intéressante aux résultats psychoacoustiques disponibles [2,7].

4. IMPLÉMENTATION DE BANC DE FILTRE GAMMACHIRP

La paramétrisation des signaux s'opère par un banc de filtre calculé à l'aide de la fonction gammachirp. Dans notre application, on utilise un banc de 32 filtres caractérisés par 32 réponses impulsionnelles gammachirp, où la fréquence central de chaque filtre gammachirp a une largeur de bande ERB et le banc couvre la bande 50-8000 Hz, Le signal de parole est segmente par une fenêtre de Hamming de largeur 25 ms. Chaque section du filtre gammachirp se compose de deux chemins, le premier chemin est le filtrage et le deuxième est l'estimation de la puissance du signal dans chaque sous bande. Le chemin de filtrage est un filtre gammatone d'ordre 4 suivi de la fonction d'asymétrie pour réaliser le filtre gammachirp final. Dans le deuxième chemin on calcule la puissance de signal dans chaque sous bande et c en utilisant l'expression suivante : $c=3.38+0.107P_s$.

La sortie du banc de filtre gammachirp est soumit à une opération de filtrage par la courbe d'égale intensité. La figure 2 donne la démarche utilisée pour déterminer le banc de filtre gammachirp, La figure 3 montre la réponse impulsionnelle de 32 filtres gammachirp, couvrant la bande de 50-8000 Hertz, après filtrage par la courbe d'égale intensité. La figue 4 donne les caractéristiques du banc de filtre gammachirp.

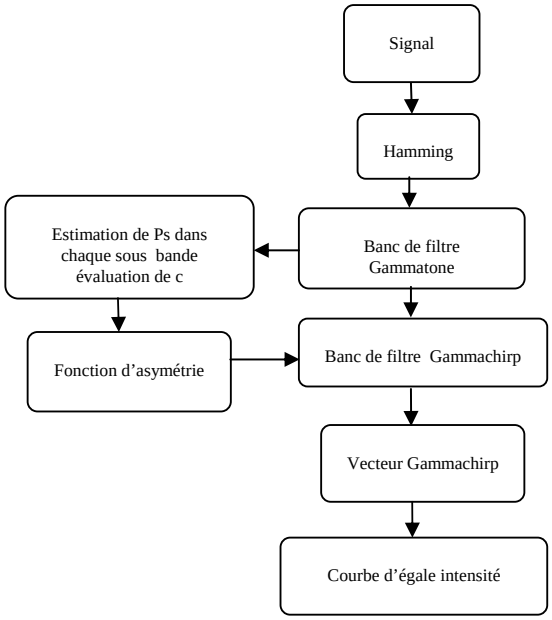


Figure 2: Bloc diagramme du filtre gammachirp.

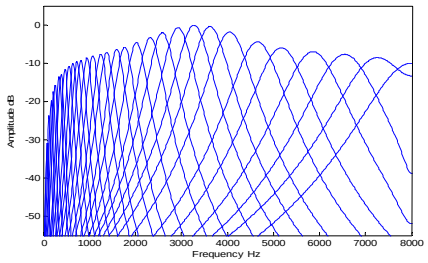


Figure 3: Exemple de banc de filtre gammachirp après filtrage par courbe d'égale intensité.

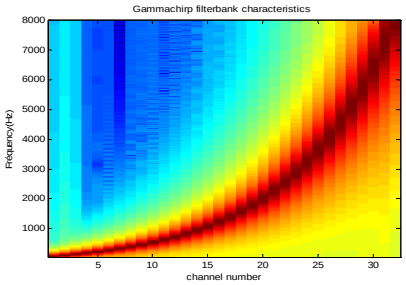


Figure 4: Caractéristique du banc de filtre gammachirp.

5. ANALYSE DE LA PAROLE BASEE SUR UN BANC DE FILTRE GAMMACHIRP

Les coefficients physiologiques du banc de filtre gammachirp peuvent être utilisé autant que les coefficients de paramétrisation du signal de parole, dans ce cas l'énergie dans chaque filtre est calculée par estimation du module de la transformée de Fourier discrète (DFT) du signal en la multipliant par le filtre gammachirp correspondant. De cette étape nous avons expérimenté deux options : La première option est GammaChirp Cepstral (GC-Cept) qui consiste à estimer les coefficients cepstraux par la transformée de cosinus discrète (DCT), ce qui décorele le signal de parole, en réduisant le nombre de coefficients d'analyse à 12 coefficients cela étant nécessaire pour le traitement du HMM. La deuxième option est GammaChirp-PLP (GC-PLP) qui est réalisée de sorte à suivre les étapes du PLP. La figure 5 donne Les différentes démarches de GC-Cept et GC-PLP.

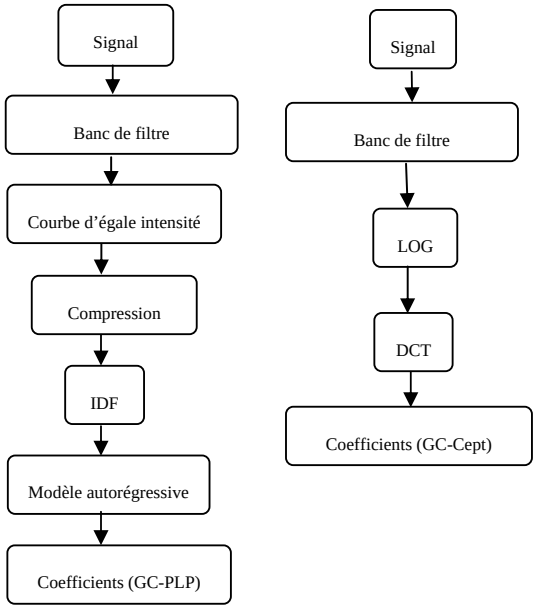


Figure 4: Parametrisation basée sur un banc de filtre gammachirp

6. EVALUATION

Pour évaluer le banc de filtre auditif gammachirp, nous avons comparé différentes techniques de paramétrisation standards (MFCC et PLP) avec les paramétrisations proposées (GC-Cept et GC-PLP) dans le cadre de la reconnaissance de mots isolés. Les 12 premiers coefficients auxquels on a ajouté l'énergie (soit 13 coefficients), les delta et les delta-delta (soit 39 coefficients). L'évaluation porte sur un corpus issu de la base TIMIT composé de 6132 mots pour la phase d'apprentissage, soit 21 mots répètes 292 fois par 18 hommes et 18 femmes réparties uniformément sur 8 dialectes américains. Pour la phase de reconnaissance nous avons utilisés 2201 mots, soit 21 mots répétés 104 fois par 13 hommes et 13 femmes réparties uniformément sur 8 dialectes américains. En ce qui concerne le modèle de Markov Cachés, HMM [9], nous avons utilisés une matrice de taille 5x5 pour l'ensemble des probabilités de transition.

Les probabilités d'émission sont de type multi gaussienne, définies uniquement par les vecteurs moyens de dimension 12, les matrices de covariance et

les poids associés à chaque gaussienne de dimension 12. Les résultats de chaque type de paramètrisation sont représentés, tout d'abord, dans un état brut, ensuite avec l'énergie, delta et delta-delta. Les tableaux 1, 2, 3 et 4 donnent les résultats associés aux taux de reconnaissances des différentes techniques de paramètrisation.
Nous définissions les paramètres si dessous.

N: Le nombre total de mots à reconnaître.
D: Le nombre de mots non pris.
S: Le nombre de mots non reconnus.
H: Le nombre de mots reconnus.
%: Le taux obtenu en pourcentage,

Table 1: Taux de reconnaissance obtenu par les techniques de paramétrisation dans leur état brut.

	%	N	H	S	D
MFCC	91.00	2201	2003	198	0
PLP	91.78	2201	2020	181	0
GC-PLP	94. 86	2201	2175	124	0
GC-Cept	91.86	2001	2025	186	0

Table 2 : Taux de reconnaissance obtenu par les techniques de paramétrisation combinées avec l'énergie.

	%	N	H	S	D
MFCC_e	93.14	2201	2050	151	0
PLP_e	94.14	2201	2072	129	0
GC-PLP_e	95.72	2201	2126	75	0
GC-Cept_e	95.20	2201	2112	89	0

Table 3 : Taux de reconnaissance obtenu par les techniques de paramétrisation combinées avec l'énergie et delta.

	%	N	H	S	D
MFCC_e_d	98.36	2201	2165	36	0
PLP_e_d	98.41	2201	2166	35	0
GC-PLP_e_d	98.94	2201	2184	17	0
GC-Cept_e_d	98.53	2201	2174	27	0

Table 4 : Taux de reconnaissance obtenu par les techniques de paramètrisation combinées avec l'énergie, delta et delta-delta.

	%	N	H	S	D
MFCC_e_d_a	98.64	2201	2171	30	0
PLP_e_d_a	99.05	2201	2180	21	0
GC-PLP_e_d_a	99.56	2201	2193	8	0
GC-Cept _e_d_a	99.10	2201	2182	19	0

7. CONCLUSION

Dans ce papier, nous avons présenté deux techniques de paramétrisations du signal de parole qui tient compte des caractéristiques fréquentielles de l'oreille, basée sur un banc de filtres dont les réponses impulsionnelles sont celles des fonctions gammachirp. Les paramétrisations implémentées ont montré leurs performances avec le système de reconnaissance automatique de la parole HTK basé sur le Modèles de Markov Cachés au vu d'une reconnaissance de mots isolés. Nous observons que les résultats les moins bons sont ceux obtenus avec la modélisation de base et les meilleurs sont ceux obtenus avec la modélisation avec banc de filtre gammachirp.

BIBLIOGRAPHIE

[1] K. Ouni. Contribution à l'analyse du signal vocal en utilisant des connaissances sur la perception auditive et représentation temps fréquence en multirésolution des signaux de parole .Thèse de Doctorat*, ENIT, 2003*.

[2] T. Irino, R. D. Patterson. Temporal asymmetry in the auditory system. *J. Acoust. Soc. Am*. 99(4): 2316-2331, April, 1997.

[3] T. Irino, D. Patterson. A time-domain, level-dependent auditory filter: the gammachirp. *J. Acoust Soc. Am.* 101(1): 412-419, January, 1997.

[4] T. Irino et M. Unoki. An analysis auditory filterbank based on an IIR implementation of the gammachirp. *J. Acoust. Soc Japan.* 20(6): 397-406, November, 1999.

[5] T. Irino, R. D. Patterson. A compressive gammachirp auditory filter for both physiological and psychophysical data. *J. Acoust Soc. Am.* 109(5): 2008-2022, may 2001.

[6] J. O. Smith III, J.S. Abel. Bark and ERB bilinear transforms, *IEEE Tran. On speech and Audio Processing*, Vol. 7, No. 6, November 1999.

[7] R. D. Patterson, I. Nimmo-Smith. Off-frequency listening and auditory-filter asymmetry, *J. Acoust. Soc. Am.,* Vol. 67, No. 1, pp. 229-245, 1980.

[8] Irino, T. and Unoki, M. (1998). A time-varying, analysis/synthesis auditory filterbank using the gammachirp. *IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-98)*, 3653-3656.

[9] H. Hermansky. Perceptual lineair predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* Vol. 87, No. 4,pp. 1738-1752., April 1990.

[10] B.R. Glasberg, B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research,* 47103-198, 1990.

[11] H. Hermansky, J. C. Junqua, Optimization of Perceptually Based ASR Front-Ends. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York*, April 11-14, 1988, paper S 5.10,pp.219-222.

Identification perceptive d'accents étrangers en français

Bianca Vieru-Dimulescu & Philippe Boula de Mareüil

LIMSI-CNRS
BP 133 – 91403 Orsay CEDEX, France
Tél.: ++33 (0)1 69 85 80 70 – Fax : ++33 (0)1 69 85 80 88
Courriel : bianca@limsi.fr ; mareuil@limsi.fr

ABSTRACT

A perceptual experiment was designed to determine to what extent naïve French listeners are able to identify foreign accents in French: Arabic, English, German, Italian, Portuguese and Spanish. They succeed in recognising the speaker's mother tongue in more than 50% of cases (while rating their degree of accentedness as average). They perform best with Arabic speakers and worst with Portuguese speakers. The Spanish and Italian accents on the one hand and the English and German accents on the other hand are the most mistaken ones. Phonetic analyses were conducted; clustering and scaling techniques were applied to the results, and were related to the listeners' reactions that were recorded during the test. Emphasis was laid on differences in the vowel realisation (especially concerning the phoneme /y/).

1. INTRODUCTION

Peut-on reconnaître l'origine d'une personne à partir d'un échantillon de parole ? Dans certains cas cela est possible : des expériences récentes l'ont confirmé sur des accents régionaux en anglais [1] et en français [2]. On sait moins dans quelle mesure des auditeurs naïfs se montrent capables d'identifier des accents étrangers en français. Le but de cet article est de combler ce manque. Des études préliminaires [3] [4] ont suggéré que le timbre des voyelles permet d'identifier l'origine d'un accent en français mieux que la prosodie (du moins le rythme). De même, la plupart des études en la matière se concentrent sur le segmental (*i.e.* la chaîne de phonèmes, en particulier vocaliques) [5] [6]. C'est cet aspect que nous développons ici.

Dans la perspective de mener des tests perceptifs, nous avons porté notre intérêt sur les accents avec lesquels des Français sont le plus susceptibles d'être familiers : allemand, anglais, arabe, espagnol, italien, portugais (plus néerlandais, uniquement conservé pour illustrer quelle réaction on peut avoir face à un accent étranger). Le choix de ces langues a été établi en recoupant des statistiques sur l'immigration et le tourisme en France. Nous avons enregistré une quarantaine de locuteurs de ces différentes langues maternelles, à la fois en lecture et en parole spontanée.

La section suivante présente le protocole et les résultats d'une expérience d'identification perceptive des 6 accents étudiés. Et la section 3, avant de conclure, examine certains traits phonétiques relevés par les auditeurs, en matière de timbre vocalique. L'analyse acoustique s'appuie sur de la lecture, des éléments de comparaison avec la parole spontanée étant fournis.

2. EXPÉRIENCE PERCEPTIVE

2.1. Locuteurs et corpus

Parmi les locuteurs d'origine allemande, anglaise, arabe, espagnole, italienne ou portugaise dont les enregistrements ont été collectés, 36 ont été utilisés pour l'expérience proprement dite (6 par langue), 6 autres (1 de chaque langue) et une locutrice néerlandaise pour une phase de familiarisation. Dans le test, on avait autant d'hommes que de femmes. Les locuteurs étaient européens ou venaient, pour les Arabes, de différents pays du Maghreb — mais une étude antérieure a montré la difficulté à discriminer les différentes origines possibles, algérienne, marocaine ou tunisienne, de locuteurs parlant français [3]. Aucun hispanophone n'était catalan ni latino-américain. L'âge moyen des locuteurs (tous étudiants) était de 24 ans, leur durée de résidence en France (dans la région parisienne) était en moyenne de 15 mois et l'âge auquel ils avaient commencé à apprendre le français était en moyenne de 17 ans.

Les locuteurs ont lu, entre autres, le texte du projet « Phonologie du Français Contemporain » (PFC) [7], et parlé librement pendant quelque 5 minutes en situation de face à face avec l'expérimentateur. Des éléments de comparaison nous sont ainsi disponibles, à travers les données du projet PFC dont nous avons repris et étendu le protocole — les locuteurs étaient également enregistrés dans leur langue maternelle.

Pour le test perceptif, un échantillon de parole spontanée, d'une dizaine de secondes, a été retenu pour tous nos locuteurs, selon les critères suivants : absence de référence culturelle et d'erreur morpho-syntaxique typique d'une origine donnée, pas trop d'hésitations et cohérence thématique de l'énoncé. Les stimuli sélectionnés ont ainsi fait l'objet d'une expérience perceptive, décrite dans ce qui suit.

2.2. Protocole et tâche des sujets

Le test se déroulait dans une chambre isolée, les auditeurs écoutaient les stimuli à travers des enceintes avec un niveau d'écoute confortable — préalablement égalisé à l'aide du logiciel Goldwave (http:// www.goldwave.com). Les stimuli, au format Wave, étaient échantillonnés à 22,05 kHz, 16 bits, mono. Le test était réalisé à travers une interface conviviale [4] [8], qui permet entre autres choses d'entrer des informations sur la familiarité avec les différents accents et de saisir les réponses.

Les sujets étaient avertis qu'ils seraient amenés à porter des jugements sur des échantillons de parole non native

en français. Ils écoutaient pour commencer un stimulus et les commentaires qu'une collègue avait faits sur la prononciation qui se trouvait être celle d'un accent néerlandais. Ensuite, en guise de familiarisation avec les accents étudiés, les sujets écoutaient un extrait de parole spontanée illustrant chacun des accents allemand, anglais, arabe, espagnol, italien et portugais. À chaque fois étaient indiqués l'origine et le degré d'accent évalué par les auteurs : de très faible à très fort. Une bonne image de la diversité des accents était ainsi fournie par ces locuteurs qui, bien sûr, n'étaient pas utilisés par la suite.

Lors de la phase suivante, le test proprement dit, les auditeurs écoutaient 36 stimuli présentés dans un ordre aléatoire différent pour chaque auditeur. Comme lors de la phase de familiarisation, chaque stimulus pouvait être réécouté, arrêté au milieu ou repris à partir d'un certain point ; mais il était impossible de revenir en arrière une fois passé au stimulus suivant. Les auditeurs devaient attribuer un degré d'accent au stimulus en question, sur une échelle continue graduée de 0 à 5. Les degrés proposés étaient paraphrasés de la façon suivante : (0) pas d'accent, (1) petit accent, (2) accent modéré, (3) assez fort accent, (4) fort accent, (5) très fort accent.

Les auditeurs, munis d'un microphone, étaient invités à réagir verbalement à l'écoute du stimulus (en l'imitant voire en le caricaturant) ou à écrire leurs commentaires dans une fenêtre de texte. Ces données étaient enregistrées stimulus par stimulus, et les consignes suggéraient simplement de préciser quels traits non natifs dans la prononciation et l'intonation du locuteur leur semblaient marquants. Les sujets devaient déterminer la langue maternelle du locuteur ayant produit le stimulus, avant de traiter un autre extrait. Le choix était forcé (sans distracteur ni classe rejet) parmi allemand, anglais, arabe, espagnol, italien et portugais. En cas d'hésitation entre deux réponses, les sujets avaient droit à une seconde chance : une option leur était facultativement laissée à cet effet. Mais elle a été très peu utilisée (seulement 30 fois pour l'ensemble des auditeurs), nous ne l'avons donc pas analysée.

2.3. Auditeurs

Le test perceptif a été soumis à 25 auditeurs de région parisienne, de langue maternelle française, sans problèmes d'audition connus et membres d'un laboratoire d'informatique (le LIMSI). Ils n'étaient pas rémunérés pour cette tâche. La table 1 reporte, entre parenthèses, le nombre d'auditeurs qui se disaient avant le test capables de reconnaître tel ou tel accent en français.

Table 1 : nombre d'auditeurs s'estimant capables de reconnaître tel ou tel accent en français (entre parenthèses) et rapportant un niveau faible, moyen ou bon dans les langues correspondantes.

Niveau	allemand (18)	anglais (20)	arabe (20)	espagnol (12)	italien (3)	portugais (6)
faible	13	0	23	14	24	24
moyen	9	5	2	10	1	0
bon	3	20	0	1	0	1

Ces chiffres ne vont pas de pair avec le niveau des auditeurs dans les langues correspondantes, également consigné : par exemple, 23 sujets sur 25 déclaraient qu'ils n'avaient pas ou que peu de connaissances en arabe, alors que 20 d'entre eux se sentaient capables de reconnaître un accent arabe en français.

2.3. Résultats

Des jugements des auditeurs, il ressort que le degré d'accent de nos locuteurs est moyen (2,66 sur 5) : 2,18 pour les Allemands ; 3,00 pour les Anglais ; 2,37 pour les Arabes ; 2,94 pour les Espagnols ; 3,07 pour les Italiens ; 2,39 pour les Portugais. Les résultats du test d'identification montrent que nos auditeurs arrivent bien à distinguer les accents étrangers présentés (*cf.* table 2). Le taux global d'identification correcte (52,2 %) est très supérieur au hasard (16,7 %). Des tests de khi-deux révèlent que pour chaque langue on est significativement au-dessus du seuil de hasard [$ddl = 3$; $p < 0{,}01$]. Et pour chaque origine linguistique la réponse majoritaire est la bonne — en gras, dans la diagonale de la table 2. Il en va de même pour 27 locuteurs sur les 36 présentés, qui sont bien identifiés.

Table 2 : matrice de confusion pour 36 stimuli et 25 auditeurs (% par rapport à 6×25 = 150 réponses).

origine \ réponse	allemand	anglais	arabe	espagnol	italien	portugais
allemand	**63,3**	14,7	6,0	3,3	4,7	8,0
anglais	28,0	**48,7**	8,7	8,7	2,7	3,3
arabe	6,0	1,3	**77,3**	2,0	5,3	8,0
espagnol	2,7	2,7	5,3	**58,7**	19,3	11,3
italien	5,3	3,3	7,3	34,0	**40,0**	10,0
portugais	17,3	8,0	16,7	20,7	12,0	**25,3**

Les locuteurs dont l'origine a le mieux été reconnue sont les Arabes, principale communauté immigrée vivant en France. Les moins bien reconnus sont les Portugais, dont les phonèmes sont « proches » du français, ce qui peut expliquer leur accent peu marqué. En outre, le stéréotype chuintant qui est souvent associé à l'accent portugais est loin de la réalité. Entre ces cas extrêmes, les confusions les plus fréquentes sont dans l'ordre italien/espagnol — ce qui est corroboré par des études antérieures sur l'accent espagnol en italien et l'accent italien en espagnol [8] — et anglais/allemand. Des techniques d'échelonnement multidimensionnel (*scaling*) et de *clustering* permettent de visualiser ce fait : elles rassemblent Italiens et Espagnols dans un même groupe, Anglais et Allemands dans un autre groupe, et fait apparaître Arabes et Portugais dans deux groupes à part (*cf.* figure 1, qui montre une sorte de distance perceptive entre les différents accents).

Des analyses de variance (ANOVA) ont également été conduites sur les réponses comptées comme correctes (1) ou fausses (0) avec le facteur aléatoire Sujet et les deux facteurs intra-sujet Familiarité (avec l'accent) et Degré d'accent. Selon que les auditeurs se sont majoritairement déclarés capables de reconnaître l'accent en français (comme dans le cas des Allemands, Anglais et Arabes) ou non (comme dans le cas des Espagnols, Italiens, Portugais), deux groupes de Familiarité ont été distingués. En ce qui concerne le Degré d'accent, les locuteurs ont été séparés en trois groupes équilibrés, moyennant les évaluations des auditeurs. Les ANOVA montrent un effet majeur de la Familiarité [$F(1,24) =$

56,5 ; $p < 0{,}01$] et du Degré d'accent des locuteurs [$F(2,48) = 21{,}4$; $p < 0{,}01$]. On a également une interaction entre les deux [$F(2,48) = 4{,}1$; $p = 0{,}02$]. Malgré cet effet global du Degré d'accent, on peut souligner qu'entre les groupes de locuteurs les mieux et les moins bien identifiés, la différence de degré d'accent entre les Arabes (2,37) et les Portugais (2,39) n'est pas significative d'après un test de Student.

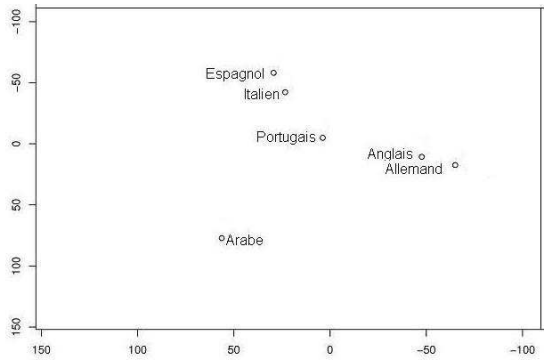


Figure 1 : résultat de l'échelonnement multi-dimensionnel **—** algorithme classique prenant en entrée une matrice de dissimilitude qui peut être calculée à partir des distances (ici euclidiennes) entre lignes de la matrice de confusion.

Les auditeurs ont déclaré s'être appuyés sur des indices dont on peut vérifier la pertinence par des mesures acoustiques. Parmi les indices segmentaux, nos auditeurs ont relevé : le *r* « roulé » qui leur évoquait des pays du Sud ; [i] à la place de /e/ dans le cas des locuteurs de langue maternelle arabe ; *yé* à la place de *je*, [v] à la place de /b/ et [s] à la place de /z/ pour les Espagnols ; [z] à la place de /s/ pour les Allemands ; [u] à la place de /y/ ou l'inverse, ainsi qu'une mauvaise réalisation des nasales, sans rapprochement avec une origine particulière, mais plutôt signe d'un accent étranger en général. D'où les analyses acoustiques suivantes, avec le tracé de triangles vocaliques des locuteurs — les études sur les consonnes et la prosodie sont en cours.

3. ANALYSES PHONÉTIQUES

Les mesures présentées dans cette section ont été faites sur le texte PFC (400 mots), lu par les 36 locuteurs utilisés dans le test de perception. Ce matériau commun, moins restreint que nos échantillons de parole spontanée, facilite l'analyse et se prête mieux aux comparaisons entre locuteurs. Même si en contrepartie, il reflète moins bien le vernaculaire (la façon naturelle de parler), ce texte permet la comparaison directe, sur le même corpus, avec des natifs français.

Grâce à l'alignement automatique dérivé du système de reconnaissance de la parole du LIMSI [9], le texte PFC a été segmenté en phonèmes en utilisant des modèles acoustiques indépendants du contexte pour le français standard. De la même façon ont été segmentés les textes lus par 6 locuteurs normands ou vendéens (3 hommes, 3 femmes) parmi les plus jeunes de ceux qui ont été étudiés dans [2]. Ces locuteurs n'avaient « pas d'accent », ou plus précisément leur degré d'accent avait été estimé à moins de 1 sur 5 par 25 auditeurs de région parisienne suivant un protocole très proche de celui de la présente étude.

Sur cette base ainsi segmentée, des études acoustiques ont été menées, facilitées par le traitement automatique de la parole. Un script a été écrit pour le logiciel PRAAT (http://www.fon.hum.uva.nl/praat/) afin d'extraire les fréquences des formants en différents points des voyelles — le texte PFC en comprend plus de 500 par locuteur. Comme la méthode de segmentation est automatique, des filtres ont été prévus (adaptés à chaque voyelle, hommes et femmes distingués) pour écarter les valeurs aberrantes à un horizon tolérant de ±500 Hz en moyenne par rapport à des valeurs de référence [10]. Seulement 4 % des phonèmes ont ainsi été rejetés. Les valeurs des premiers formants ont ensuite été normalisées à l'aide de diverses procédures décrites par [11]. Utilisant la normalisation de Nearey, les triangles vocaliques correspondant aux différents accents (ou origines linguistiques) sont donnés figure 2, où les valeurs des formants au tiers, à la moitié et aux deux tiers des voyelles sont moyennées. On peut noter que les triangles des locuteurs anglais et allemands sont plus réduits que les autres — ce qu'on peut relier à la réduction vocalique que connaissent leurs langues d'origine. L'antériorisation du /u/ anglais est également remarquable, de même que le fait que parmi les /e/, le plus proche des [i] est celui des Arabes.

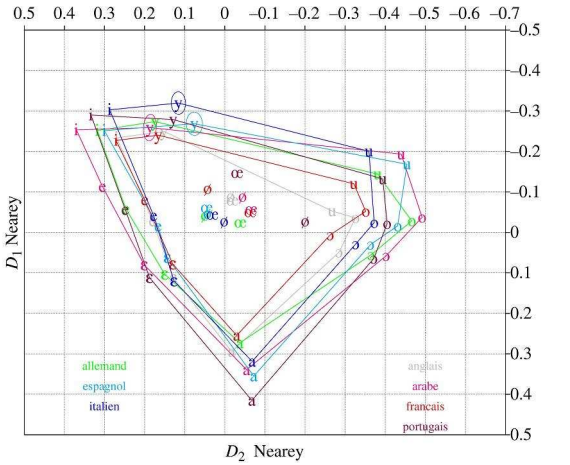


Figure 2 : triangles vocaliques (après filtrage des probables erreurs de mesure et normalisation de Nearey) correspondant au français natif ou parlé avec différents accents, pour le texte PFC. De gauche à droite sont entourés d'ellipses les /y/ des Arabes, des Italiens et des Espagnols.

La réalisation du /y/ français est particulièrement différente entre les locuteurs espagnols ou italiens notamment (chez qui elle est plus proche du [u]) et les locuteurs arabes (chez qui elle tend vers le [i]). Les uns privilégient le trait [+antérieur], les autres le trait [+arrondi]. Ce phénomène souvent caricaturé est connu [12], même si on n'en a pas d'explication . On peut le retrouver dans des transcriptions ludiques telles que *tou m'as toué* ou bien *Itats-Inis*. Il est bien mis en évidence par scaling ou clustering à partir d'une caractérisation de chaque origine par les coordonnées moyennes de son /y/ dans le plan F1/F2. Une représentation graphique

sous forme de dendrogramme en est fournie figure 3. En introduisant le troisième formant (F3), en utilisant un autre algorithme (agglomératif plutôt que divisif) et quelle que soit la distance utilisée (euclidienne ou Manhattan), on obtient des résultats très semblables. On retrouve la même tendance sur les phrases spontanées présentées aux auditeurs, que nous avons transcrites, alignées et analysées comme le texte lu.

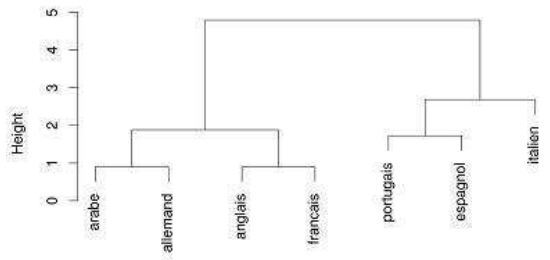


Figure 3 : dendrogramme issu du clustering divisif utilisant une distance euclidienne, à partir d'une caractérisation dans l'espace acoustique de la prononciation du /y/ en français standard et dans chaque accent.

En ne représentant plus les moyennes par langue d'origine mais par locuteur, on obtient par le même algorithme divisif utilisant une distance euclidienne un dendrogramme où les locuteurs arabes d'une part et les locuteurs italiens et espagnols, d'autre part, sont assez bien regroupés : 5 Arabes sur 6 du côté gauche, 5 Italiens et 4 Espagnols sur 12 du côté droit de l'arbre. Ce /y/ pourrait donc être un bon indice discriminant.

4. CONCLUSION

Ainsi, il a été possible de mener une étude à base de perception, d'analyse de données et de traitement automatique de la parole, sur les accents allemand, anglais, arabe, espagnol, italien et portugais en français. Ces accents étrangers sont ceux avec lesquels nous avons le plus de chance d'être exposés. Quelque quarante locuteurs ont été enregistrés, parlant français avec un accent jugé moyen. Leur origine a été bien identifiée par des auditeurs français natifs à partir d'échantillons de parole — à 52 %, soit 10 % de plus que 6 accents régionaux en français, étudiés dans une tâche similaire [2]. Ces auditeurs ont indiqué de façon opportune les indices acoustiques qui leur paraissaient saillants. Pour hiérarchiser ces derniers, nous poursuivons nos mesures de VOT (*Voice Onset Time*) pour les consonnes occlusives, et devons lier les paramètres rythmiques [4] avec le ratio de durée des syllabes accentués / non accentuées. Il restera à comparer les résultats avec les productions en langue maternelle de nos locuteurs.

Plusieurs perspectives s'ouvrent pour le traitement automatique : ajout de variantes liées aux accents étrangers (ex. /y/ [u], avec les modèles acoustiques français) ; alignements avec les modèles acoustiques des langues d'origine (ex. [ɹ] anglais), en vue d'une identification automatique des accents et d'une amélioration des scores de reconnaissance de la parole non native. La synthèse de la parole mériterait également d'être mise à profit : mieux que des imitateurs humains qui ont trop tendance à renforcer certains traits, cet outil est un bon instrument de simulation. Enfin, nous espérons que cette étude pourra être utile pour l'apprentissage et l'enseignement du français langue étrangère.

5. REMERCIEMENTS

Nous sommes reconnaissants à Martine Adda-Decker, Cécile Woehrling et Cédric Gendrot pour la mise à disposition, le lancement ou l'adaptation de scripts notamment exploitant l'alignement automatique.

6. BIBLIOGRAPHIE

[1] C.G. Clopper & D.B. Pisoni. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32: 111-140, 2004.

[2] C. Woehrling & P. Boula de Mareüil. Identification d'accents régionaux en français : perception et catégorisation. *Bulletin PFC*, 6 : 89-103, 2006.

[3] P. Boula de Mareüil, B. Brahimi & C. Gendrot, Role of segmental and suprasegmental cues in the perception of Magrebian-accented French. *Interspeech-ICSLP*, Jeju, 2004.

[4] B. Vieru-Dimulescu, P. Boula de Mareüil & M. Adda-Decker. Identification perceptive d'accents étrangers en français : premiers résultats, *10^es Journée PFC*, 2006.

[5] B. Lauret. *Aspects de Phonétique Expérimentale Contrastive : « l'accent» anglo-américain en français.* Thèse de doctorat, Université Paris III, 1998.

[6] J.E. Flege, C. Schirru & I.R.A. MacKay, Interaction between the native and second language phonetic subsystems, *Speech Communication*, 40(4) : 467-491, 2003.

[7] E. Delais-Roussarie & J. Durand (éd.), *Corpus et variation en phonologie du français. Méthodes et analyses.* Presses Universitaires du Mirail, Toulouse, 2003.

[8] B. Vieru-Dimulescu & P. Boula de Mareuil, Contribution of prosody to the perception of a foreign accent: A study based on Spanish/Italian modified speech. *ISCA Workshop on Plasticity in Speech Perception*, Londres, 2005 (pp. 66-69).

[9] M. Adda-Decker, P. Boula de Mareüil, G. Adda & L. Lamel. Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, 46 (2): 119-139, 2005.

[10] C. Gendrot & M. Adda-Decker, Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *Interspeech*, Lisbonne, 2005 (pp. 2453-2456).

[11] P.M. Adank. *Vowel normalization : a perceptual-acoustic study of Dutch vowels.* Thèse de doctorat, Radboud University Nijmegen, 2003.

[12] B.L. Rochet, Perception and Production of Second-Language Speech Sounds by Adults, *in* W. Strange (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, York Press, York, 1995 (pp. 379-410).

Vers un inventaire ordonné des configurations manuelles de la Langue des Signes Française

Leïla Boutora

Laboratoire SFL, UMR 7023 – Université Paris 8
2, rue de la Liberté – 93526 Saint-Denis Cedex
leila.boutora@neuf.fr

ABSTRACT

This article deals with French Sign Language (FSL), and particularly with the question of the definition of its minimal units (in realisation and perception). The general aim of this work is to know if we can make strict equivalence between phonems, the minimal units of realisation of vocalic languages (VL), and the minimal units of sign languages (SL). We make a focus on the status of handshapes in FSL in the lexicon and on the problem we are faced with the definition of their inventory in « phonetic » terms.

1. INTRODUCTION

L'expression des langues signées (LS) repose sur un canal de communication différent de celui des langues vocales (LV) puisqu'il est visuel-gestuel *versus* audio-vocal. De ce fait, une LS exploite de manière pertinente, c'est-à-dire linguistique, les trois dimensions spatiales en plus de deux dimensions temporelles – séquentielle et simultanée – au sein d'une grammaire « spatiale ». Cela donne aux locuteurs la possibilité d'exprimer de manière simultanée des informations grammaticales et lexicales par le jeu des gestes manuels (décomposables en paramètres de configuration, orientation, emplacement et mouvement manuels) combinés à des éléments non manuels tels que la direction du regard (informations syntaxiques), l'expression du visage (mimiques faciales adjectivales , aspectuelles…) ou encore la tête et le corps (fonction phatique, frontière thématique).

La question du statut des unités minimales des LS ne connaît pas de consensus, particulièrement pour les éléments du paramètre *configuration*, les uns les assimilant aux phonèmes des LV, les autres refusant des parallèles trop simplistes ou trop rapides. Nous souhaitons apporter des éléments de réponse en nous fondant sur la perception des locuteurs de la langue des signes. Cependant, avant de pouvoir réaliser de tels tests, et pour la Langue des Signes Française (LSF) du moins, le problème de la définition d'un inventaire unifié de ses configurations manuelles en termes « phonétiques » se pose. Nous tentons actuellement d'élaborer un tel inventaire qui prenne en compte dans sa classification les caractéristiques articulatoires et morphémiques des configurations de la LSF.

2. UNE DOUBLE ARTICULATION EN LS ?

2.1. Affirmer que les LS sont des langues

En 1960, Stokoe [22], linguiste américain, montrera dans une perspective structuraliste que l'ASL (American Sign Language) est doublement articulée et pour ce faire mettra en évidence les paramètres manuels de formation du signe (ou *kinème*) dont les éléments paradigmatiques sont posés comme équivalents de phonèmes (*chérèmes*). Cette position évacue de fait la question de l'iconicité dans la langue qui n'est *a priori* pas compatible avec la double articulation. S'ensuivra pendant quarante ans un nombre important de travaux – avec chacun leur intérêt et leurs limites – qui chercheront à montrer les parallèles que l'on peut observer entre les LS et les LV (Klima & Bellugi [17] pour les traits distinctifs, Miller [19] pour la syllabe et bien d'autres). Les travaux qui questionnent l'iconicité des langues des signes et/ou remettent en cause le transfert de modèles des LV vers les LS sont encore peu répandus outre-atlantique (proposition de Stokoe [23] d'une *Semantic phonology*, ou d'Uyechi [24] d'une *Visual phonology*).

2.2. Prendre en compte les caractéristiques du canal visuel-gestuel

En France, la description de la LSF a été entreprise avec un décalage de vingt ans par rapport à l'ASL. Une partie importante des études la concernant ont porté sur la description des structures iconiques et sur la formalisation des relations actancielles, ou encore l'expression de la temporalité. On aura un bon aperçu des travaux les plus récents dans [1] et [2]. L'étude approfondie des éléments sublexicaux de la LSF, quand à elle, reste un thème délicat puisque la question de la double articulation et de l'iconicité en LS qui semblait être réglée pour certains, est réactualisée par les questionnements de plusieurs chercheurs qui proposent différentes solutions à cet épineux problème, soit en redéfinissant la notion de phonème comme unité purement « oppositive » qu'elle soit porteuse de sens ou non (Bouvet [6]), soit en accordant une place secondaire à l'iconicité sans pour autant la reléguer au domaine extra-linguistique (Jouison [16]), ou encore en proposant de distinguer le traitement des différents paramètres (Millet [20]). Cuxac [10] sera le seul à

accorder une place centrale à l'iconicité en postulant une bifurcation de visées dans son système, donnant naissance à deux catégories de productions qui interagissent au sein d'un même énoncé : 1) des structures volontairement iconiques très productives (les SGI : Structures de Grande Iconicité) et 2) les signes lexicalisés en partie iconiques, mais sans intention du locuteur. Mais certaines questions restent en suspens : Comment qualifier d'équivalents de phonèmes des éléments qui sont dans une grande partie du lexique porteurs de sens ? Que reste-t-il de la double articulation dans les langues des signes dans ces conditions ?

3. DES « PHONÈMES » DANS LES LS ?

Le phénomène de perception catégorielle mis en évidence pour les LV par Liberman et al. [18] a permis de rendre compte de la réalité psychologique du phonème en tant qu'entité catégorielle. De récents travaux américains (Emmorey et al. [14]) ont porté sur des tests de CP de deux paramètres manuels de l'ASL : la configuration et le mouvement. Les résultats de ces travaux ont montré que le paramètre de configuration était perçu catégoriellement par les locuteurs de l'ASL tandis que le mouvement était perçu de manière continue.

3.1. Perception catégorielle de la configuration manuelle en ASL

Les tests comparant la perception de sourds signeurs et d'entendants non signeurs ont porté sur des paires de configurations « phonologiquement » contrastives, i.e. considérées comme des équivalents de phonèmes, et sur des paires de type « allophonique ». Le continuum entre deux configurations extrêmes a été généré avec un outil de génération 3D, le logiciel *Poser*©. Deux hypothèses sont testées :

1) Si les locuteurs sourds ont une perception catégorielle et pas les entendants non signeurs, alors la perception catégorielle est à la base du langage indépendamment de la modalité – on sait par ailleurs que la perception catégorielle est présente dans des domaines autres que linguistiques ;

2) Il n'y a CP que pour les unités linguistiquement contrastives.

Les résultats indiquent que seuls les sourds signeurs montrent un pic de discrimination qui correspond à la frontière des catégories phonémiques et seulement pour les couples de phonèmes.

Pourtant, une autre étude (Emmorey & Herzig [13]), qui ne portait pas sur la perception catégorielle mais sur le statut linguistique ou non d'éléments possédant des caractéristiques iconiques, montre que les locuteurs sourds ont une perception graduelle des configurations de « classificateurs », autrement dit, des configurations qui reprennent la forme d'un actant cataphoriquement ou anaphoriquement, ou encore qui composent les structures iconiques définies dans Cuxac [10] consistant en la reprise de forme ou de taille d'un actant de l'énoncé. Cette catégorie de configurations n'a pas été l'objet de test de CP dans la première étude citée [14]. Cette étude ne prend pas en compte la dimension morphémique, voire iconique des unités, alors que cela paraît primordial pour l'interprétation des résultats : à savoir si, selon leur nature, toutes les configurations sont perçues de la même manière par les locuteurs et les non locuteurs.

3.2. Perception catégorielle de la configuration manuelle en LSF ?

Nous avons donc entrepris d'effectuer le même type de tests sur les configurations de la LSF. Les problèmes méthodologiques observés plus haut nous conduisent toutefois à être particulièrement attentifs au type de configurations à soumettre au test de CP pour la LSF. En effet, nous émettons les hypothèses que 1) une part importante du lexique possède des configurations de type morphémique, comme l'avance Cuxac dans [11] et [12], et que 2) les configurations morphémiques et non morphémiques ne sont pas nécessairement perçues de la même manière par les locuteurs et les non locuteurs de la LSF. Si l'analyse du lexique décrite dans Boutora [5] permet de vérifier la première hypothèse, l'élaboration des tests de CP devra alors tenir compte de cette donnée pour vérifier la seconde. En outre, cette analyse nous permettra de caractériser et d'estimer les proportions des différentes catégories de configurations présentes dans les signes lexicalisés, à savoir les configurations de SGI, celles de la dactylologie (alphabet manuel utilisé dans l'épellation de noms propres dont on retrouve les configurations sous forme d'initialisation dans les signes), des chiffres, ou de toute autre catégorie mise au jour.

Cependant, un certain nombre de problèmes se posent toujours à nous. Nous ne disposons pas d'un inventaire unifié et reconnu des configurations de la LSF, ni d'un système de notation efficace, et encore moins d'une classification régulière. Enfin, d'un inventaire à l'autre, les variations au sein de continuum ne sont pas appréhendées de la même manière.

4. RECOUPEMENT DES INVENTAIRES

4.1. Description des inventaires

Afin d'être en mesure d'élaborer des tests de perception catégorielle de manière rigoureuse, nous avons entrepris de définir un inventaire des configurations manuelles de la LSF en recoupant neuf inventaires qui comportent de 30 à 139 configurations.

4.2.1. Cuxac [10]

Cuxac [10] propose un inventaire non fermé de 39 configurations relevées dans les SGI de la LSF.

Certaines s'inscrivent dans un continuum, et toutes apparaissent aussi dans les signes lexicalisés. Elles ont toutes une ou plusieurs valeurs morphémiques le plus souvent prototypiques.

Tableau 1 : Exemples de valeurs prototypiques de configurations manuelles de la LSF dans les SGI

	épaisseur en fonction de l'écartement pouce / autres doigts
	saisie de formes fines et minces (cuiller, clé, carte…)
	Fin d'un déploiement de forme allongée se terminant en pointe, envisagée en volume plein

4.2.2. IVT [15]

IVT (1998) présente 61 configurations dans sa grammaire ; dans le dictionnaire, parmi les 4500 entrées classées par configuration, 4 sont dites « divers », i.e. dont les configurations ne sont pas répertoriées dans la grammaire (celle de FENETRE, H ou « cornes », qui est en fait bien répertoriée dans la grammaire ; celle de SCOUT : W dactylologique mais doigts serrés, issue de la gestuelle co-verbale ou celle de LIT variante ouverte de H à deux mains ; et enfin une variante de M plutôt productive, répertoriée dans la grammaire, mais sans la pliure du poignet qui est due à une contrainte articulatoire).

4.2.3. Bouvet [6]

A partir de l'ancienne édition d'IVT qui propose 50 configurations, Bouvet retire 44 unités articulatoires dans un premier temps, puis 39 configurations de base et 16 variantes, soit 55 unités. Sa démarche est fondée sur une classification articulatoire indiquant l'identification des doigts déployés et le mode de leur déploiement.

4.2.4. Bonucci [4]

Bonucci (1998), dont la démarche clairement « phonologique » a été « d'éliminer toute information soit prédictible soit non catégorielle », dénombre 30 configurations « cardinales » à partir de la classification de Bouvet (1992).

4.2.5. Brugeille [8]

Cet inventaire de 50 configurations ordonnées selon une logique de déploiement progressif de la main (proposées sous forme vidéo) présente l'intérêt particulier d'être proposé par un locuteur natif de la LSF qui effectue des recherches sur sa langue dans le cadre de l'association Iris.

4.2.6. Bonnal [3]

Cet inventaire de 44 configurations est né de l'élaboration d'un dictionnaire historique de la LSF (projet en cours) et rend compte d'un état de langue du 19[e] siècle. Il repose sur une classification numérique qui suit le déploiement des doigts lorsque l'on compte de 0 (poing fermé) à 5 (les 4 doigts et le pouce tendus et écartés), puis propose une déclinaison par traits articulatoires (doigts fléchi/tendus, serrés/écartés, pince ouverte/fermé/ronde…).

4.2.7. Braffort [7]

Dans une analyse articulatoire visant à élaborer un modèle de reconnaissance de gestes, Braffort dégage 55 configurations statiques sur un total de 139 comprenant en outre des configurations dynamiques, i.e. composées d'une configuration de début et d'une configuration de fin qui sont liées par un mouvement d'ouverture ou de fermeture de la main mettant en jeu les mêmes doigts avec un arrangement qui évolue progressivement entre le début et la fin du signe. Les deux formes manuelles peuvent s'opposer par un ou plusieurs traits. Pour passer de 5p à bec5, on passe de doigts écartés à doigts serrés et de pince ouverte à pince fermée. Cet inventaire comporte en outre les configurations « bille », « boule » et « 2p- » qu'aucun des inventaires « standard » ne proposent. On retrouve les deux premières dans les SGI. La troisième appartient à la « famille » – pour ne pas dire « catégorie » prématurément – des « becs d'oiseau » mais qui marque une position intermédiaire supplémentaire du continuum qui se déploie entre « bec ouvert » (2p) et « bec fermé » (bec2), par rapport aux autres inventaires qui ne proposent qu'une seule position intermédiaire pour ce couple ouvert-fermé.

Cette observation rejoint celle effectuée dans Cuxac [10] et suscite inévitablement une question concernant la légitimité du choix d'une, de deux ou de plusieurs configurations intermédiaires parmi un continuum de possibilités articulatoires. Nous noterons avec intérêt le résultat du test de CP qui portera sur ce type de paire de configurations.

4.2.8. Companys [9]

Cet inventaire ne prétend pas à l'exhaustivité puisqu'il propose « quelques » configurations au nombre de 46 dans un *mode d'emploi* de mise en œuvre des SGI dégagées par Cuxac [10]. Il a tout de même retenu notre attention pour les raisons suivantes : il est le seul à proposer la configuration « H ouvert » qui apparaît dans au moins une occurrence du dictionnaire IVT (signe LIT). Il confirme la nécessité de la configuration « pouce ouvert » (signe AMI) présent aussi chez Braffort [7]. Il nous montre des variantes de configurations utilisées à Angers (A et I). Enfin, il est proposé par des locuteurs de la LSF.

4.2.9. Nancy [21]

Une équipe de Nancy qui élabore actuellement un lexique spécialisé de LSF en ligne propose 48 configurations sur son site Internet. Les configurations listées constituent une entrée possible pour la

recherche de signes. Nous n'avons pas relevé de logique d'organisation particulière.

Inventaire provisoire

A partir des neuf inventaires décrits ici, nous avons dégagé par recoupement 77 configurations statiques. A partir de cet ensemble, nous allons identifier les configurations de base et leurs variantes, articulatoires et morphémiques. Nous avons retenu la logique de classification numérique proposée par Bonnal, affinée au niveau de la description en traits articulatoires.

Il s'agit ensuite de déterminer quelle est la part des configurations potentiellement morphémiques, et de préciser dans quelles proportions chaque configuration possède cette valeur dans le lexique d'IVT.

BIBLIOGRAPHIE

[1] A.M. Bertonneau et G. Dal (eds). Linguistique de la LSF : recherches actuelles. Actes du colloque de Villeneuve d'Ascq, 23-24 septembre 2003. In *Silexicales 4*, 2004.

[2] M. Blondel et L. Tuller (eds). Langage et surdité, *Recherches Linguistiques de Vincennes* 29, 2000

[3] F. Bonnal. *Sémiogenèse de la langue des signes française : étude critique des signes attestés sur support papier depuis le XVIII[e] siècle et nouvelles perspectives de dictionnaires*, Thèse de doctorat, Université Toulouse le Mirail, 2005.

[4] A. Bonucci. A*nalyse phonologique et indexation figurative pour une base de données d'entrées lexicales de la Langue des Signes Française*. Thèse de doctorat, Université Lyon 2, 1998.

[5] L. Boutora. Une phonologie de la LSF (Langue des signes française) ? Vers une détermination de ses unités minimales et de leur(s) statut(s). *In ADL 2005, Paris, 17 et 18 octobre 2005*, 2005.

[6] D. Bouvet. Classification articulatoire des configurations de la main dans la langue des signes française : Portée heuristique de cette classification pour la recherche des unités distinctives. *Protée,* 20-2: 23-32 et 20-3: 87-99, 1992.

[7] A. Braffort. *Reconnaissance et compréhension de gestes, application à la langue des signes*. Thèse de doctorat en informatique, Université de Paris 11 – Orsay, 1996.

[8] J.L. Brugeille, cf. site d'IRIS, section *Recherche* : http://membres.lycos.fr/iris/

[9] M. Companys (ed). *La langue des Signes Française : mode d'emploi*, Editions Monica Companys, Angers, 2003.

[10] C. Cuxac. La Langue des Signes Française : les Voies de l'Iconicité, *Faits de Langues* 15-16, Ophrys, Paris, 2000.

[11] C. Cuxac. Compositionnalité sublexicale morphémique iconique en LSF, Blondel M. et Tuller L. (eds) *Recherches Linguistiques de Vincennes* 29: 55-72, 2000.

[12] C. Cuxac. Phonétique de la LSF : une formalisation problématique, *Silexicales 4* :. 93-113, 2004.

[13] K. Emmorey and M. Herzig. Categorical versus gradient properties of classifier constructions in ASL. In Emmoery (ed). *Perspectives on classifier constructions in sign languages*. Lawrence Erlbaum Associates Inc, Mahwah, NJ, pp. 221-246, 2003.

[14] K. Emmorey, S. McCullough, D. Brentari. Categorical perception in American Sign Language, *Language and Cognitive Processes*, 18-1: 21-45, 2003.

[15] M. Girod (dir.). *La Langue des Signes*. Tome 1 Histoire et grammaire, Tomes 2 et 3 Dictionnaire bilingue LSF/Français, 2ème édition, Editions IVT, Paris, 1998.

[16] P. Jouison. Iconicité et double articulation dans la langue des signes. In S. Quertinmont et F. Loncke (eds), *Etudes européennes en Langues des Signes*, Edirsa, Bruxelles, pp. 75-107, 1989.

[17] E. Klima and U. Bellugi (eds). *The Signs of Language*, Harvard University Press, Cambridge London, 1979.

[18] A.M. Liberman, K.S. Harris, H.S. Hoffman, B.C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Pscychology*, 54: 358-368, 1957.

[19] C. Miller. *La phonologie dynamique du mouvement en langue des signes québécoise*, Saint-Laurent, Québec, Fides, 2000.

[20] A. Millet. Réflexions sur le statut du mouvement dans les langues gestuelles : aspects lexicaux et syntaxiques. *Lidil* 15: 11-30, 1997.

[21] Nancy, Lexique LSF : www.lsf.univ-nancy2.fr

[22] W. C. Stokoe. *Sign Language Structure*. Studies in Linguistics. Occasional Papers n° 8. Buffalo, NY, University of Buffalo Press, 1960.

[23] W.C. Stokoe. Semantic phonology. *Sign Language Studies*, Silver Springs, Maryland, pp. 107-114, 1991.

[24] L. Uyechi. *The Geometry of Visual Phonology*, Stanford, California, CSLI Publications, 1996.

A propos du trait ATR des voyelles nasales du twi

Kofi Adu Manyah

Institut de Phonétique de Strasbourg
LILPA EA 1339 Composante Parole & Cognition
Université Marc Bloch, 22 rue Descartes, BP 80010, 67084 Strasbourg, France
Tél : 33 (03)88 41 73 64 – Fax : +33 (03)88 41 73 69
Email : manyah@umb.u-strasbg.fr – http : //misha1.u-strasbg.fr/IPS

ABSTRACT

This paper investigates acoustic properties of the Twi nasal vowel counterparts of the oral vowels /i/ vs /ɪ/ and /u/ vs /ʊ/ investigated in a previous study of [+ATR] and [-ATR] vowels. Acoustic measurements are carried out to investigate for differences between vowel quality in the 2 groups. The evidence from our acoustic data, confirming results obtained for the oral vowels, is the tendency for advanced vowels [+ATR] to have lower F1 values, higher F2 and F3 values than the unadvanced vowels [-ATR].

1. INTRODUCTION

Cette recherche prolonge un précédent travail acoustique, sur le trait ATR des oppositions /i/ vs /ɪ/ et /u/ vs /ʊ/ [2]. Nous avons proposé d'examiner, dans l'étude préliminaire, quelques aspects acoustiques des contrastes vocaliques [ATR], à savoir les structures formantiques des oppositions des voyelles orales /i/ vs. /ɪ/ et /u/ vs. /ʊ/ et leurs durées. Les résultats des oppositions des voyelles nasales /ĩ/ vs. /ɪ̃/, et /ũ/ vs. /ʊ̃/ sont présentés dans cet article. Notre travail sera divisé en trois parties, d'inégale longueur. Une première partie, rappellera tout d'abord, la situation du twi par rapport aux langues akan. La deuxième abordera le corpus et la méthodologie de notre protocole expérimental. La troisième partie présentera les résultats et discussions de nos analyses.

L'harmonie vocalique ATR en akan ou twi a fait l'objet de diverses analyses [8, 9, 10, 11, 12, 13, 15, 16]. Le twi, une langue à quantité et à deux tons ponctuels, est parlé par les Asante du Ghana. La population asante parle le twi, d'où l'appellation asante twi utilisée par certains ; le twi fait partie des langues akan. Les langues akan englobent les populations du Ghana qui parlent le twi dans la région Asante et dans certaines parties des régions 'Eastern', 'Western', 'Central', 'Volta' et 'Brong Ahafo' : à savoir les Asante, les Akyem, les Kwawu, les Akuapem, les Wassa, les Twifu, les Assin, les Gomoa et les Fante. L'asante appartiendrait au sous-groupe kwa (Niger-Congo) du groupe nigéro-kordofanien, de la grande famille négro-égyptien : Obenga [14]. Pour certains linguistes, les langues akan couvrent une région ouest africaine plus vaste, s'étendant de la région Sud-Est de la Côte d'Ivoire jusqu'à la région du Volta du Ghana. En résumé, clarifions que dire akan simplement, sans d'autres précisions, impliquerait que l'on admette le choix d'interprétations diverses et que l'on ne tienne pas compte d'un paramètre majeur de notre étude, c'est-à-dire la spécificité de la langue twi des Asante, par rapport aux autres langues akan. Pour notre travail expérimental sur les propriétés acoustiques du trait ATR en twi, nous avons choisi des natifs de la région asante parlant twi.

Précisons également qu'à notre connaissance aucune étude, ou très peu d'études [15], sur le trait ATR concernant ces langues, n'a abordé la classe phonologique des nasales et les 2 groupes phonologiques bref et long.

Les 9 voyelles orales twi, dont 5 ont des homologues nasales /ĩ, ɪ̃, ã, õ et ũ /, sont divisées en deux séries : une position avancée de la racine de la langue (*Advanced Tongue Root position* [ATR]) pour un groupe de voyelles et une position moins avancée pour l'autre. Le premier groupe [+ATR], est réalisé avec une position abaissée du larynx, donnant un pharynx moins contracté, et le deuxième groupe [-ATR], est produit avec une position élevée du larynx, aboutissant à une cavité pharyngale plus contractée [9, 10, 13].

Groupe 1 : [+ATR]			**Groupe 2 : [-ATR]**	
i	**u**		**ɪ**	**ʊ**
e	**o**		**ɛ**	**ɔ**
a		⟷	**a**	

La voyelle /a/ est la seule exception en ce sens qu'elle peut figurer dans les 2 cas (cf. schéma ci-dessus). Bien que la voyelle /a/ appartienne phonétiquement au groupe [-ATR], elle n'a pas d'équivalent dans le groupe [+ATR] avec laquelle elle contraste.

Il existe des cas progressifs mais l'harmonie vocalique est essentiellement régressive en twi. Dans le cas de syllabes successives, plus précisément dans une séquence polysyllabique, lorsque la deuxième syllabe contient une voyelle fermée, la voyelle ouverte de la première syllabe est remplacée par la voyelle fermée correspondante.

Les 2 types d'assimilation, simple dans le cas des mots dissyllabiques, et complexe, dans le cas des mots trisyllabiques, ont été aussi évoqués dans l'étude précédente [2].

Nous proposons d'examiner, dans cette deuxième étude, quelques aspects acoustiques des contrastes vocaliques [ATR], à savoir les structures formantiques des oppositions nasales /ĩ/ *vs.* /ɪ̃/ et /ũ/ *vs.* /ʊ̃/ et leurs durées.

2. MÉTHODE

Les locuteurs étaient deux adultes masculins de langue maternelle twi, sans antécédent pathologique du conduit vocal et possédant une audition normale. Le corpus était constitué des mots monosyllabiques, comportant des oppositions brèves et longues, dans des environnements consonantiques C1VC2 où C1 est /p/, /k/ ou /t/ et C2 est

/k/. Les paires minimales ont été insérées dans une phrase porteuse. Les phrases ont été ensuite transcrites sur des fiches et présentées en ordre aléatoire. Chaque phrase a été répétée au moins 10 fois par les deux locuteurs. Les enregistrements acoustiques ont été réalisés, en vitesse d'élocution normale, dans une chambre insonorisée, stockés numériquement et analysés. Dans un premier temps, des mesures de durées ont été prélevées pour la voyelle cible et la consonne post-vocalique C2, le /k/, ce qui nous permet d'obtenir 3 intervalles : la durée vocalique; la durée consonantique et la durée totale V+C. Nous avons procédé à un traitement statistique classique à l'ensemble des données. Ainsi, nous disposons de moyennes, écart-type et *t* de Student ($p \leq 0.01$). Dans un deuxième temps, nous avons procédé à une analyse de formants par le biais de l'éditeur de signal Praat. Pour chaque séquence de contraste phonémique, 3 mesures de 4 valeurs formantiques (F1, F2, F3, F4) ont été prélevées : au début, au milieu et à la fin des réalisations vocaliques. À partir des valeurs brutes ainsi obtenues, nous avons pu calculer une moyenne pour les valeurs formantiques de chaque réalisation vocalique et ensuite calculer une moyenne pour les 10 répétitions pour chaque locuteur.

3. RÉSULTATS ET DISCUSSION

Les figures ci-dessous (figure 1, figure 2, figure 3 et figure 4) montrent la tendance générale de nos résultats les plus significatifs, à savoir, les valeurs formantiques des voyelles pour chaque contraste. Comme pour nos résultats des voyelles orales, l'analyse des valeurs absolues montre que les valeurs formantiques donnent des indications quant à la réalisation des oppositions [+ATR] / [-ATR] des voyelles nasales.

3.1 Opposition /ĩ/ vs. /ɪ̃/

Pour le premier locuteur, les résultats des analyses acoustiques montrent que les voyelles nasales avancées [+ATR] longues ont des valeurs formantiques F1 moins élevées que leurs homologues non-avancées [-ATR]. Ces résultats vont dans le sens de ceux que nous avons obtenus pour les voyelles orales correspondantes. En revanche, dans la catégorie des brèves les voyelles nasales avancées ont des valeurs formantiques F1 plus élevées que leurs homologues non-avancées.

Pour le deuxième locuteur, les résultats montrent que les voyelles nasales avancées [+ATR] ont des valeurs formantiques F1 moins élevées que leurs homologues non-avancées [-ATR] dans les deux classes. Signalons cependant que ces différences ne sont pas toujours très nettes, restant parfois à un état de tendance. Dans la catégorie des longues, la voyelle nasale avancée /ĩ/ a une valeur moyenne pour F1 de 255 Hz (61 Hz). Son homologue non-avancée /ɪ̃/ possède une valeur moyenne de 401 Hz (26 Hz) [les écarts types sont entre parenthèses]. Les résultats de ce locuteur vont dans le sens des résultats obtenus pour les voyelles orales.

À l'inverse des valeurs formantiques du F1, les valeurs de F2 des voyelles avancées sont plus élevées que celles des voyelles non-avancées. Encore une fois, il serait plus prudent de parler de tendances dans certains cas.

Pour le premier locuteur, les valeurs formantiques de F3 dans cette catégorie sont plus hautes pour la voyelle avancée /ĩ/ brève que la voyelle non-avancée /ɪ̃/. Les valeurs de la voyelle non-avancée sont plus élevées que celles de la voyelle avancée dans la série des longues. Pour le deuxième sujet, contrairement au premier sujet, c'est plutôt le F3 de la voyelle non-avancée /ɪ̃/ qui est plus haut dans la catégorie des brèves. Dans le groupe des longues, les valeurs de F3 de la voyelle avancée sont plus élevées que celles de la voyelle non-avancée.

Nos données temporelles [1], [4], [5], [6], montrent que les durées vocaliques des deux groupes sont comparables pour les deux locuteurs (cf. table 1).

Table 1 : durées des oppositions /ĩ/ *vs.* /ɪ̃/ pour les deux locuteurs (ms)

Locuteur 1 +ATR	-ATR
ĩ bref = 113 (14)	ɪ̃ bref = 1O7 (16)
ĩ long = 279 (23)	ɪ̃ long = 308 (34)
Locuteur 2 +ATR	-ATR
ĩ bref = 87 (09)	ɪ̃ bref = 78 (10)
ĩ long = 195 (25)	ɪ̃ long = 186 (17)

3.2 Opposition /ũ/ vs. /õ/

L'analyse des valeurs spectrales dans ce contexte présente une tendance plus cohérente que celle observée pour l'opposition précédente. En effet, les voyelles avancées [+ATR] ont des valeurs formantiques F1 moins élevées que leurs homologues non-avancées [-ATR], dans les deux catégories. Les figures 1 et 2 montrent que la voyelle avancée /ũ/ brève affiche une valeur moyenne pour F1 de 232 Hz (37 Hz) pour le locuteur 1. Son homologue non-avancée /õ/ a une valeur moyenne de 330 Hz (57 Hz). Dans la catégorie des longues la voyelle avancée /ũ/ révèle une valeur moyenne pour F1 de 239 Hz (25 HZ). Son homologue non-avancée /õ/ affiche une valeur moyenne de 281 Hz (66 Hz). Les valeurs correspondantes pour le locuteur 2 sont de 251 Hz (37 Hz) pour /ũ/ brève et 352 Hz (83 Hz) pour /õ/ brève. Dans la catégorie des longues, les résultats obtenus pour le deuxième sujet indiquent une moyenne de 259 Hz (37 Hz) et 386 Hz (63 Hz) pour /ũ/ et /õ/ respectivement (cf. figure 4).

Dans ce contexte, et contrairement aux valeurs formantiques F1, les valeurs de F2 des voyelles avancées ont tendance à être plus élevées que celles de leurs homologues non-avancées. La voyelle avancée /ũ/ brève indique une valeur moyenne pour F2 de 1667 Hz (497 Hz) pour le premier locuteur (cf. figure 1). Son homologue non-avancée montre une valeur moyenne pour F2 de 1262 Hz (86 Hz). Dans la catégorie des longues, la voyelle avancée /ũ/ a une valeur moyenne de F2 de 1552 Hz (371 Hz) pour le sujet 1. Son homologue non-avancée /õ/ affiche une valeur moyenne de 1278 Hz (129 Hz). Les valeurs correspondantes pour le locuteur 2

sont de 1431 Hz (303 Hz) et 1391 Hz (177 Hz) respectivement (cf. figure 4).

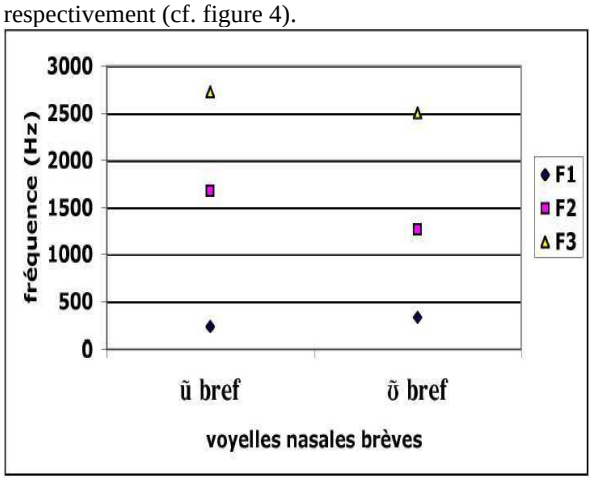


Figure 1 : Valeurs formantiques des oppositions +ATR /u/ nasal bref *vs* -ATR /ʊ/ nasal bref pour le locuteur 1

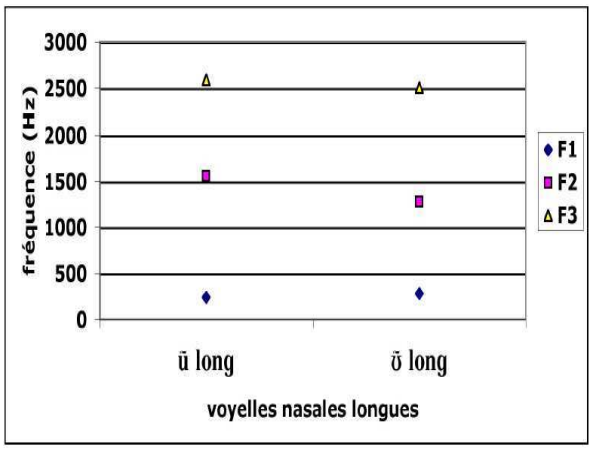


Figure 2 : Valeurs formantiques des oppositions +ATR /u/ nasal long *vs* -ATR /ʊ/ nasal long pour le locuteur 1

Les valeurs formantiques de F3 dans cette catégorie montrent des valeurs plus hautes pour la voyelle avancée que pour son homologue non-avancée quel que soit le sujet. Les figures 1 et 2 montrent que, s'agissant du premier locuteur, la voyelle non-avancée /õ/ possède une valeur moyenne inférieure de 2500 Hz (62 Hz) et une valeur moyenne supérieure de 2724 Hz (492 Hz) pour la voyelle avancée /ũ/ brève. Les valeurs correspondantes pour le locuteur 2 sont de 2353 Hz (53 Hz) et 2431 Hz (193 Hz) pour /õ/ et /ũ/ respectivement.

Dans la série des longues, les valeurs formantiques de F3 confirment la tendance observée dans la série des brèves. La voyelle avancée /ũ/ longue affiche une valeur moyenne de 2591 Hz (300 Hz) et la voyelle non-avancée /õ/ affiche une valeur moyenne de 2505 Hz (112 Hz) pour le premier sujet.

Les valeurs correspondantes pour le deuxième sujet sont de 2476 Hz (194 Hz) pour la voyelle avancée et 2394 Hz (81 Hz) pour la voyelle non-avancée (cf. figure 4).

Dans cette catégorie, comme dans la catégorie précédente, les durées des deux groupes de voyelles sont comparables quel que soit le locuteur (cf. table 2).

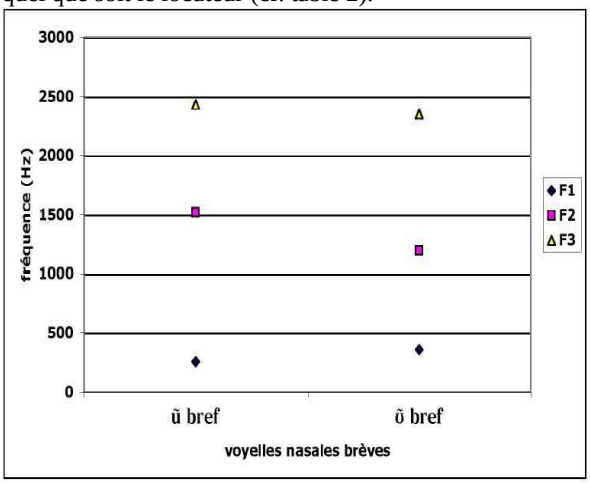


Figure 3 : Valeurs formantiques des oppositions +ATR /u/ nasal bref *vs* -ATR /ʊ/ nasal bref pour le locuteur 2

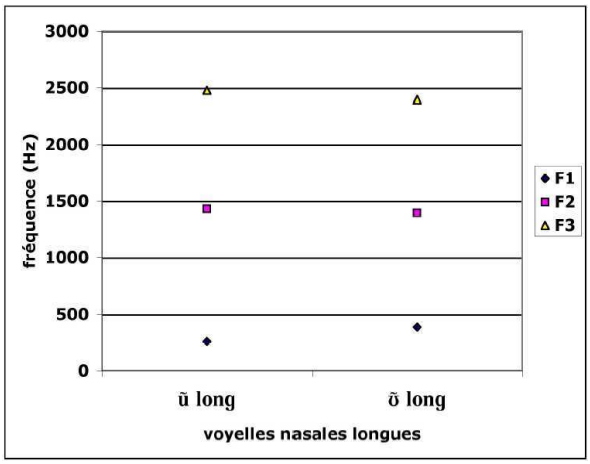


Figure 4 : Valeurs formantiques des oppositions +ATR /u/ nasal long *vs* -ATR /ʊ/ nasal long pour le locuteur 2

Table 2 : durées des oppositions /ũ/ *vs*. /õ/ pour les deux locuteurs (ms)

Locuteur 1 +ATR	-ATR
ũ bref = 146 (24)	õ bref = 127 (14)
ũ long = 293 (36)	õ long = 300 (36)
Locuteur 2 +ATR	-ATR
ũ bref = 87 (13)	õ bref = 106 (15)
ũ long = 198 (38)	õ long = 179 (22)

4. CONCLUSION

L'analyse des données acoustiques des voyelles nasales a montré que les valeurs formantiques des voyelles pouvaient permettre la distinction des classes, même si la distinction reste au niveau de tendances dans de nombreux cas. Contrairement aux résultats de l'opposition /ɨ/ *vs.* /ɪ/, les résultats des voyelles nasales /ũ/ et /õ/ vont dans le sens des observations faites sur les

voyelles orales, dont les valeurs moyennes de F2 et F3 sont relativement plus élevées que celles des voyelles nasales.

Il semble tout de même que c'est l'effet global des différences de valeurs formantiques qui reflète mieux les contrastes. Sachant que F3 donne des indications sur la formation d'une cavité labiale, sur la projection et à l'arrondissement des lèvres, ou la protrusion, que F2 indique la position avant/arrière de la masse linguale, i.e. le lieu d'articulation et F1 renseigne sur le degré d'aperture c'est-à-dire, la distance entre la voûte palatine et le dos de la langue et sur la position du larynx : Calliope [7], nous pouvons faire les observations suivantes.

La catégorie [+ATR] a, en général, des valeurs formantiques de F1 plus basses (ce qui pourrait être révélateur d'une augmentation de la cavité pharyngale par rapport à la classe [-ATR]), un deuxième formant (F2) plus haut et un F3 et un F4 plus hauts que ceux de la classe [-ATR].

La comparaison de la structure formantique des deux classes indique une structure relativement plus compacte pour la classe [-ATR] par rapport à son homologue [+ATR], avec un F1 élevé et un F2 relativement plus bas pour le [-ATR] par rapport au [+ATR] (F1 plus faible et F2 plus élevé).

La comparaison de la structure du F3 des deux classes pourrait indiquer une structure relativement plus arrondie, plus protruse pour la classe [+ATR] par rapport à son homologue [-ATR], en ce qui concerne la classe des voyelles postérieures /ũ/ et /õ/. Cela pourrait contribuer à clarifier des observations sur le trait ATR en twi. D'abord, dans une étude sur 1 sujet Asante et 3 sujets Akyem, Lindau [13] suggère que les différences entre les F3 sont négligeables, et conclut que F1, ayant plus d'intensité, est le corrélat acoustique le plus important de l'harmonie vocalique c'est-à-dire, de la position de la racine de la langue. Ensuite, dans leur présentation, Ladefoged & Maddieson [11] procèdent à une estimation de positions des lèvres. On peut penser que dans ces études l'accent était mis sur les changements intervenant à l'intérieur de la cavité buccale plutôt que sur ceux liés à une modification de la cavité labiale.

Les résultats acoustiques ont montré que, pour les deux locuteurs, les durées vocaliques sont comparables dans le groupe [+ATR] et le groupe [-ATR].

La suite de ce travail sur le trait ATR des voyelles du twi consistera à vérifier les formants pour les contrastes /e/ *vs*. /ɛ/, /o/ *vs*. /ɔ/. Nous procéderons aussi à une étude articulatoire de ce phénomène phonologique qui nous livrera de véritables informations, quant à la configuration du conduit vocal lors de la production de ces contrastes.

Remerciements : Je remercie R. Sock pour ses remarques et suggestions. Cette recherche a été partiellement financée par le Programme ACI–TTT 2003-2006 du Ministère de la Recherche et des Nouvelles Technologies attribué à l'Institut de Phonétique de Strasbourg, LILPA EA 1339 Composante Parole & Cognition, Université Marc Bloch.

BIBLIOGRAPHIE

[1] K. Adu Manyah. Quantité et qualité vocaliques en twi : le cas des voyelles nasales. *XVèmes Rencontres Linguistiques en Pays Rhénan*, Université Marc Bloch Strasbourg 2, pp. 9-30, 2005

[2] K. Adu Manyah. Harmonie vocalique et ATR en twi : aperçu phonologique et étude acoustique préliminaire. *XXVèmes Journées d'Étude sur la Parole (JEP) de l'AFCP*, Fès, Maroc, 2004.

[3] K. Adu Manyah. *Étude contrastive du système phonologique en akan (twi) et du système phonologique en français en vue d'une application didactique*. Lille, Diffusion ANRT, 2004.

[4] K. Adu Manyah. Vowel quantity contrasts in Twi. *15th International Congress of Phonetic Sciences*, Barcelona, pp. 3185-3188, 2003.

[5] K. Adu Manyah. *Introduction à la phonétique et à la phonologie africaines. Les sons de tous les jours : le cas akan (TWI)*. Paris, l'Harmattan, 2002.

[6] K. Adu Manyah & R. Sock. La quantité vocalique en twi. Quelques considérations phonologiques et analyses acoustiques préliminaires. *XXIVèmes Journées d'Étude sur la Parole (JEP) de l'AFCP*, Nancy, pp. 41-44, 2002.

[7] Calliope. *La parole et son traitement automatique*. Paris, Masson, Collection Technique et Scientifique des Communications, 1989.

[8] G. N. Clements. Akan vowel harmony: a nonlinear analysis. *Harvard Journal of Phonology*, No. 2 pp. 1O8-177, 1981.

[9] S. Hess. Assimilatory effects in a vowel harmony system: an acoustic analysis of advanced tongue root in Akan. *Journal of Phonetics*, No. 20 pp. 475-492, 1992.

[10] S. Hess. Acoustic characteristics of the vowel harmony feature and vowel raising in Akan. *UCLA Working Papers in Phonetics*, No. 68 pp. 58-72, 1987.

[11] P. Ladefoged & I. Maddieson. *The Sounds of the World's Language*. Oxford UK, Cambridge USA, Blackwell Publishers, 1996.

[12] M. Lindau-Webb. Tongue mechanisms in Akan and Luo. *UCLA Working Papers in Phonetics*, No. 68 pp. 46-57, 1987.

[13] M. Lindau. The feature expanded. *Journal of Phonetics*, No. 7 pp. 163-176, 1979.

[14] T. Obenga. *Origine Commune de l'Égyptien Ancien, du Copte et des Langues Négro-Africaines Modernes. Introduction à la linguistique historique africaine*, Paris, L'Harmattan, 1993.

[15] C. Painter. Pitch control and pharynx width in Twi: an electromyographic study. *Phonetica*, Vol. 33, pp. 334-352, 1987.

[16] J. M. Stewart. Tongue Root Position in Akan Vowel Harmony. *Phonetica*, Vol. 16, No. 4 pp. 185-204, 1967.

Variation, coup de glotte et glottalisation en persan

Assadi Sh. S.

Laboratoire de Phonétique et Phonologie (UMR 7018) CNRS / Sorbonne Nouvelle
19, Rue des Bernardins. 75005 Paris. France.

Tél : 0l43263780 Fax : 0144430573
suassadi@yahoo.fr

ABSTRACT

Glottalization phenomena are produced with much variation across individual speakers and between speakers. Data analysed here is consisted of 228 isolated words, 3 short texts and 2 dialogues (15 minutes), read by two native speakers of Persian. The paper describes the inter- and intra-speaker variability in the realization of glottalization , the effect of context and pitch accent.

1. INTRODUCTION

Les phénomènes de glottalisation sont réalisés avec beaucoup de variations intra et inter-locuteurs, et cette observation a été faite dans de nombreuses langues. Ils peuvent être produits par une fermeture totale de la glotte ou par des vibrations irrégulières des cordes vocales Ladefoged et Maddieson [11]. Cette irrégularité peut se manifester au niveau de la forme, de l'amplitude ou de la durée des périodes vibratoires. La glottalisation peut perturber la détection de la fréquence du fondamental et des traitements subséquents dans les systèmes automatiques. Elle peut aussi fournir des informations pertinentes sur les plans segmental et prosodique.

Les phénomènes de glottalisation peuvent avoir des fonctions diverses. Nous traitons ici les deux fonctions suivantes en persan : (i) « l'occlusive glottale » avant la voyelle initiale de mot, qui est marqueur de frontière, position dans laquelle elle n'est pas distinctive ; (ii) « l'occlusive glottale » en position médiane et finale de mot, elle se trouve dans les mots d'emprunt arabes et est considérée comme un phonème. Il est à noter que du point de vue historique « l'occlusive glottale » du persan correspond à la pharyngale sonore et au coup de glotte arabe.

La présente recherche fait partie d'une analyse phonétique détaillée des phénomènes de glottalisation dans divers corpus : mots et phrases isolés, textes en différents débits et parole spontanée de plusieurs locuteurs de différentes régions d'Iran (Assadi [1]). La présente analyse se limitera à la prononciation de Téhéran.

Les études phonétiques antérieures sur les phénomènes de glottalisation en persan sont principalement basées sur les aspects auditifs. En ce qui concerne la présence d'un coup de glotte avant la voyelle initiale de mot, les auteurs sont partagés. Des ambiguï tés demeurent, par ailleurs, sur la réalisation de « l'occlusive glottale » au milieu et fin des mots d'emprunt arabes. Ainsi, Gaprivdashvili & Giunashvili [8], en s'inspirant de l'arabe, considèrent « l'occlusive glottale » du persan comme une pharyngale sonore alors que Kavitskaya [10] l'examine comme une approximante.

Nos recherches ont montré beaucoup de variations en fonction de l'origine du locuteur, du style et du débit de la parole (Assadi [1]).

Dans ce travail, nous étudions (i) les variations inter-locuteurs, (ii) les variations en fonction du corpus (mots isolés, textes et parole spontanée) et (iii) l'effet de l'accent d'insistance sur le coup de glotte et /ou la glottalisation de la voyelle initiale de mot.

2. METHODE

2.1. Corpus

Le corpus est constitué de :

- 228 mots isolés (inscrits sur des cartes) et montrés aux locuteurs à un intervalle de 5 seconde,
- 3 textes courts, lus le plus rapidement possible,
- 15mn de conversation.

2.2. Locuteurs

Deux locutrices originaires de Téhéran ont participé à cette expérience.

3. ANALYSE

L'analyse du corpus a été effectuée des points de vue acoustique et perceptif. Le terme de *glottalisation* a été attribué aux striations verticales irrégulières sur le

spectrogramme (figure 2) correspondant aux vibrations irrégulières des cordes vocales (figure 1) et le *coup de glotte* à un silence suivi d'une « explosion » (figure 3) correspondant à la fermeture complète de la glotte. Du point de vue perceptif, ces deux phénomènes sont perçus comme une qualité craquée et comme une attaque dure.

3.1. Variation selon le locuteur

Afin d'examiner les variations inter-locuteurs, 42 mots commençant tous par la voyelle [*a*] ont été examinés. Ce choix vient du fait que l'occurrence des mots commençant par la voyelle mentionnée est plus fréquente dans notre corpus que les autres mots. Le signal a été zoomé au maximum et nous avons étudié :

- la durée des périodes irrégulières au début de la voyelle ainsi que la forme et l'amplitude de celles-ci (figure 1),

- la durée totale de la voyelle initiale de mot

Le pourcentage de la glottalisation de chaque voyelle a été ultérieurement calculé. Les moyennes de la durée de la voyelle initiale de mot ainsi que sa portion glottalisée sont illustrées dans le tableau 1 pour les deux sujets.

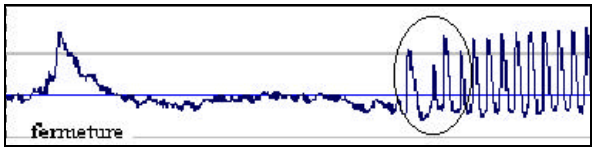


Figure 1 : Exemple des périodes irrégulières au début de la voyelle initiale de mot en persan (la figure ci-dessus illustre un signal laryngographique) .

Tableau 1 : Moyenne de la durée de la voyelle initiale, celle de la partie glottalisée ainsi que le pourcentage de la glottalisation pour chaque sujet.

	Durée (ms)		
	voyelle	portion glot.	moy glot.
sujet A	184	32	17%
sujet B	198	21	11%

En ce qui concerne l'amplitude et la forme du signal, chaque période a été comparée avec les périodes adjacentes (voir Lieberman, [15]). Les caractéristiques de chaque période ont été notées. L'irrégularité du signal peut se manifester par la forme et / ou par la durée et/ou par l'abaissement de l'amplitude. En moyenne, 17 % de la durée de la voyelle est glottalisée par le sujet A et 11 % par le sujet B. La différence entre les durée de glottalisation n'est donc pas significative. Par contre, la variation au niveau des autres caractéristiques acoustiques du signal comme l'amplitude et la forme irrégulières est importante. Cette étude confirme celle de Dilley et Shattuck-Hufnagel [3] sur la diversité des réalisations acoustiques de la glottalisation chez les locuteurs.

3.2. Variation selon le corpus

Umeda [1] indique qu'en anglais, différents corpus pourraient avoir divers degrés de glottalisation. Afin d'examiner cette hypothèse en persan, nous avons comparé l'effet du coup de glotte et glottalisation dans divers corpus (mots isolés, textes et parole spontané),

Nos données montre qu'en position initiale de mot, la voyelle peut commencer par un coup de glotte et/ou par une glottalisation. La différence entre les corpus se situe essentiellement au niveau de la réalisation du coup de glotte. Par contre, la glottalisation est la réalisation fréquente dans tous les corpus étudiés. Le coup de glotte est une sorte de renforcement par rapport à la glottalisation (Ficher-Jørgensen [3]).

En position initiale, le pourcentage de coups de glotte est moins élevé dans le texte qu'en parole spontanée. Ceci est dû à l'enchaînement des mots qui aboutit ainsi à la disparition du coup de glotte.

Le tableau 2 illustre le pourcentage des coups de glotte dans différents corpus. Les résultats montrent qu'il y a une hétérogénéité entre la distribution du coup de glotte dans les positions initiale, médiane et finale des corpus étudiés. Cependant, le test chi 2 simple indique que la différence entre les trois corpus ne reste significative qu' en position médiane et finale. La variance théorique est égale à : 55 / 2 = 27,50 et chi 2= 22,27 avec $p<0,0001$.

La figure 3 illustre un exemple de coup de glotte + glottalisation (en position médiane de mot) et la figure 2 montre un cas de glottalisation (en fin de mot).

En ce qui concerne l'étude du texte, il n'y a que 10% de coups de glotte en position médiane et finale. Dans 33% des mots étudiés, il y a une glottalisation et 13% sont marqués par un simple changement d'amplitude. Par ailleurs, on observe la disparition de l'occlusive glottale dans 44% des mots.

En persan parlé, la disparition de « l'occlusive glottale » en milieu et fin des mots d'emprunt arabes aboutit à l'allongement de la durée de la voyelle précédente. La durée compensatoire résultant de l'amuï ssement des phonèmes /ʔ/ et /h/ a déjà été

mentionnée dans la littérature persane. Nous ne traitons donc pas la durée compensatoire dans cette étude et nous nous limitons à la durée perçue (Durand [5]) et non pas à la durée mesurée. On sait que de nombreux paramètres comme les segments précédents, l'accent, la position et le nombre de phonèmes dans la syllabe peuvent influencer la durée mesurée.

Nous avons donc vu que les mots isolés sont plus susceptibles de comporter un coup de glotte que les mots en contexte.

Cependant, le coup de glotte et la glottalisation de la voyelle initiale de mot peut être influencée par différents facteurs comme la pause Umeda [13], l'accent emphatique Dilley & al. [4],. Carton & al. [2], Pierrehumbert [12] et Kohler [9] ainsi que la structure prosodique Pierrehumbert [12], Dilley & al. [4] et Fougeron [7].

A présent, nous allons examiner l'effet de l'accent sur la glottalisation de la voyelle initiale de mot.

3.3. Analyse des mots accentués vs non accentués

Afin de savoir si l'accent favorise l'occurrence du coup de glotte et/ou de la glottalisation de la voyelle initiale de mot, nous avons relevé tous les mots mis en relief et les avons comparés avec leur correspondants non accentués. Ainsi, 20 mots mono-syllabiques mis en relief ont été examinés et comparés avec 57 mots non accentués. Il faut préciser que l'accent lexical en persan est en général situé sur la syllabe finale des mots. En choisissant les mots mono-syllabiques, l'accent tombe donc sur les voyelles cibles. Le contexte précédent reste constant et se termine par une voyelle. La comparaison s'est fait donc ainsi :

- voyelle # voyelle cible (mots mono-syllabiques, mis en relief).

- voyelle # voyelle cible (mots mono-syllabiques, non mis en relief).

Tableau 2 : Pourcentage d'occurrence de coup de glotte dans divers corpus.

	mots isolés	textes	parole spontanée
Position initiale	37%	12%	29%
Position médiane et finale	35%	10%	0

Tableau 3 : Occurrence du coup de glotte et/ou de la glottalisation dans les mots accentués et non-accentués

	mots mis en relief (n=20)	**mots non-mis en relief (n=57)**
coup de glotte	12	6
glottalisation	8	20
enchaînement et autres cas	0	31
total coup de glotte et /ou glottalisation	20 (100%)	(26) 45%

Les résultats montrent un coup de glotte et/ou une glottalisation dans 100% des voyelles cibles accentuées. Alors que de telles réalisations sont observées dans 45% des mots non-accentués et dans le reste des cas, on observe soit un enchaînement qui avec la dernière syllabe du mot précédent aboutissant ainsi à la disparition de la glottalisation soit une très faible irrégularité dans l'amplitude des périodes adjacentes. Aucun mot mis en relief n'est marqué par l'enchaînement ou l'absence de la glottalisation (tableau 3).

4. CONCLUSION

Cette recherche nous a permis de montrer les variations des phénomènes de glottalisation en persan qui met en évidence la dynamique du système de production. Nous avons pu également observer certaines régularités : pour chaque locuteur, on observe une sorte de hiérarchie dans la réalisation des signaux acoustiques allant du coup de glotte à l'abaissement de l'amplitude, en passant par la glottalisation. La différence entre les trois corpus se situe au niveau de la réalisation du coup de glotte. Ce dernier indique un trait de renforcement par rapport à la glottalisation. Les différentes versions (coup de glotte et glottalisation) correspondent selon Ficher-Jørgensen [3] à différents degrés d'activité des muscles du larynx. Ainsi, différentes situations de communication et de style (Lindblom [16]) contribuent aux différents degrés de la précision articulatoire, elle-même dépendante de l'effort du locuteur (Vaissière [14]). Cette étude contribue à certains aspects universels des phénomènes de glottalisation dans les langues basés sur les gestes articulatoires.

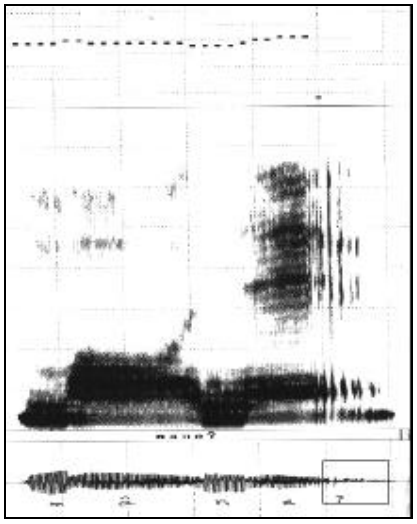


Figure 2 : Spectrogramme du mot [mɑneʔ] (empêchement). La partie encadrée illustre une glottalisation).

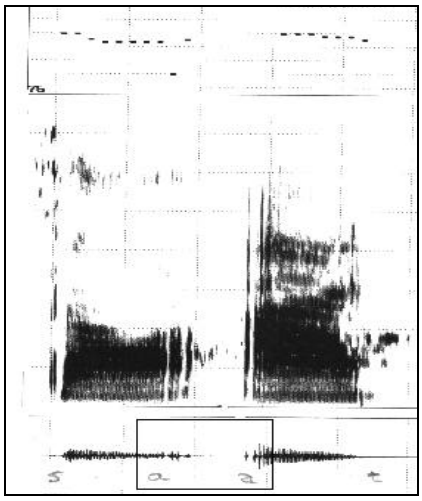


Figure 3 : Spectrogrammes du mot [sɑʔat] (horloge).

La partie encadrée montre une glottalisation+ coup de glotte+ glottalisation.

BIBLIOGRAPHIE

[1] Sh. S. Assadi. *Les phénomènes de glottalisation en persan (langue standard / langue parlée*). Thèse de doctorat. Université de la Sorbonne Nouvelle. Paris. 2003.

[2] F. Carton, D. Hirst, A. Marchal & A. Seguinot. *L'accent d'insistance / Emphatic stress*. Didier. Montréal. Paris. Bruxelles. 1977.

[3] L. Dilley and S. Shattuck-Hufnagel. Variability in glottalization of word onset vowels in American English. *Proceedings of the XIIIth international congress of phonetic sciences.* Stockholm. Vol. 4. pp : 586-589. 1995.

[4] L. Dilley, S. Shattuck-Hufnagel & M. Ostendorf. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, pp: 423-444. 1996.

[5] M. Durand. (1946). *Voyelles longues et voyelles brèves*. Klincksiek. Paris.

[6] E. Fischer-Jorgensen. Phonetic analysis of the stod in Standard Danish. *Phonetica* 46, pp. 1 –59. 1989.

[7] C. Fougeron. *Variations articulatoires en début de constituants prosodiques de différents niveaux en français.* Thèse de doctorat. Université de la Sorbonne Nouvelle. Paris. 1998.

[8] Sh. Gaprivdashvili, & J. Giunashvili, (*The Phonetics of Persian Language*. Oriental Institute of the Academy of Science. Georgie 1964.

[9] K. J. Kohler. Glottal stops and glottalization in German . Data and theory of connected speech processes. *Phonetica* 51, pp. 38 – 51, 1994.

[10] D. Kavitskaya. *Compensatory Lengthening: Phonetics, Phonology, Diachrony*. Routledge (Taylor and Francis). 2002.

[11] P. Ladefoged & I. Maddieson. *The Sounds of the World's Languages*. Oxford : Blackwell. 1996.

[12] J. Pierrehumbert. *Prosodic Effects on Glottal Allophones*. In : *Vocal Fold Physiology*. Fujimura 0. and Hirano M. Eds., pp : 39-60. Singular Publishing Group, Inc. San Diego. California. 1995.

[13] N. Umeda. Occurrence of Glottal Stops. *Jasa*. vol. 64, n. 1, pp. 88-94. 1978.

[14] J. Vaissière (2001). Changements de sons et variations synchroniques: du latin au français. *Revue Parole,* pp : 53-88.

[15] P. Lieberman. Perturbations in vocal pitch. *Journal of the Acoustica Society of America*. Vol. 33. no. 5, pp : 597-603. 1961.

[16] B. Lindblom Explaning phonetic variation: A sketch of the H & H theory. In : *Speech production and speech modelling*. Hardcastle W.J. and Marchal A. Eds, pp : 403- 439. Dordrecht : Kluwer Academic publichers. Netherlands.1990.

Influence de la distribution et des caractéristiques acoustiques sur la perception des bilingues et des monolingues

Cas du /r/ chez les guadeloupéens et chez les français

Johanne AKPOSSAN-CONFIAC

Université de Paris III – Sorbonne Nouvelle
Institut de Linguistique et de Phonétique Générales et Appliquées
Laboratoire de Phonétique et Phonologie CNRS UMR 7018
19, rue des Bernardins - 75005 Paris, FRANCE
Tél. : ++33 (0)1 43 26 37 80
Courriel : joh_akpossan@yahoo.fr

ABSTRACT

First language shapes perception and production (Troubetzkoy, 1939; Kuhl et al., 1992). This paper compares Guadeloupean with French listeners. In Guadeloupe, creole and french coexist. The bilingual Guadeloupeans are exposed to two different phonological systems where consonant /r/ has different distribution and acoustic characteristics. Perceptual, phonological and statistical analyses tend to show an influence of these both parameters on speech perception.

1. INTRODUCTION

La perception d'un individu est influencée par le système phonologique de sa langue maternelle. Et, dès 8 mois de vie, il s'opère une réorganisation de l'espace perceptif autour des phonèmes de la langue maternelle [6] et un filtre phonologique se met en place pour conditionner sa perception de telle sorte que l'enfant perçoit et produit, dès lors, des sons sur la base de la langue de son environnement [9]. Chaque langue possède son propre système phonologique et à chaque système phonologique est associé un filtre phonologique. L'individu devient alors « sourd » aux structures étrangères à sa langue. Aussi, nous interrogeons-nous sur l'influence que peut avoir une différence de caractéristiques distributionnelles et acoustiques d'un son sur la perception des individus. Nous nous sommes intéressée au /r/ des guadeloupéens (bilingues créole-français) et à celui des français (monolingues français). En Guadeloupe, créole et français coexistent dans un contexte de bilinguisme social (milieu diglossique où le créole constitue la langue de statut social inférieur). Ces deux langues ont chacune leur système phonologique. En créole, /r/ ne se trouve jamais en position finale de mot, est fricative vélaire et se réalise comme vélaire labialisée en contexte labial ce qui lui vaut d'être confondu avec la semi-consonne /w/ du créole et du français [10]. Tandis qu'en français, le /r/ est une fricative uvulaire et peut se trouver dans toutes les positions du mot.

L'admission de /r/ en français dans toutes les positions du mot, « comble » t-elle la déficience du /r/ en finale en créole dans la perception des guadeloupéens (bilingues créole-français)?

Comment le /r/ (vélaire) des guadeloupéens en contexte labial (/r/ labio-vélaire semblable au /w/ labiovélaire du créole et du français) est-il perçu par les guadeloupéens et par les français métropolitains ?

2. CORPUS ET METHODOLOGIE

2.1. Corpus

Corpus: Le corpus, enregistré en chambre sourde, est constitué de 49[1] mots monosyllabiques en français standard (cf. Annexes ; tableau 1), avec et sans /r/. Sur l'ensemble des 49 mots, 30 ont un /r/ en attaque et/ou en coda (ex : *rat* / *art* /*rare*) et les 19 autres en diffèrent par la seule absence de /r/ (ex : *a)*.

2.2. Participants

2 locuteurs: 1 guadeloupéen ; 25 ans ; niveau bac qui a vécu et grandi en Guadeloupe et est capable de tenir aussi bien une discussion en français qu'en créole et a « l'accent créole ». 1 français parisien, 27 ans, niveau bac + 4, de la région parisienne et qui ne parle pas créole. Ils ont chacun la liste de mots contenus dans le corpus (cf. Annexes ; tableau 1).

20 auditeurs: 10 guadeloupéens (5 garçons et 5 filles) âgés de 25 à 35 ans et 10 français métropolitains (4 garçons et 6 filles) âgés de 35 à 50 ans. Les guadeloupéens parlent couramment créole et français. Les français parlent français et ne sont pas familiers avec le créole.

[1] La liste de mots constituant le corpus aurait pu être plus longue et ainsi plus complète mais, elle a néanmoins permis de dégager un certain nombre de tendances.

2.3. Méthodologie de l'analyse du corpus

Nous avons effectué à partir de ce corpus 3 types d'analyses différentes : perceptive, phonologique et statistique.

2.3.1. Test de perception et analyse phonologique

Test de perception : Les 10 auditeurs guadeloupéens et les 10 auditeurs français ont écouté la liste de mots prononcés par chacun des 2 locuteurs (soit 98 mots) et ils ont pour tâche d'inscrire dans une grille ce qu'ils entendent. Il leur est permis de réécouter les stimuli autant de fois qu'ils le désirent. On obtient ainsi pour chacun des 2 groupes de 10 auditeurs un total de 980 réponses (98 mots*10 auditeurs).

Les mots consignés dans les grilles des auditeurs sont comparés à ceux de la liste de mots du corpus.

Analyse phonologique : L'analyse phonologique ne porte que sur le /r/. Les discordances entre les réponses des auditeurs et les mots du corpus sont définies en terme de suppression (ex : « art » est perçu « a »), de substitution (ex : « lard » est perçu « lave »), d'ajout (ex : « poids » est perçu « proie ») et de déplacement (ex : « foire » est perçu « froid »). Ainsi quand un auditeur perçoit, par exemple, « pause » à la place de « ose » (ajout de /p/) ou « art » au lieu de « lard » (suppression de /l/), nous estimons qu'il n'a pas été commis d'erreur. De même, si à l'écoute du mot « rare » (mot avec un /r/ à l'initiale et un autre en finale) l'auditeur perçoit « gras », cela compte pour une double faute car le /r/ initial est remplacé par /gr/ et le /r/ final est supprimé. Et enfin, nous qualifions également de substitution toute intervention de phonème qui n'appartiendrait pas au mot lu (ex : « proie » est perçu « toi » où /pr/ devient /t/). En revanche, il n'est pas question de substitution dans le cas où /pr/ devient /p/ mais plutôt d'une suppression puisque /p/ était déjà dans le mot lu.

2.3.2. Analyse statistique

A l'aide du logiciel Statview, nous exploitons les résultats des analyses perceptives et phonologiques en vue d'établir des statistiques qui permettent d'évaluer leur significativité. Le seuil de significativité est fixé à $p<0,05$.

3. ANALYSES ET RÉSULTATS

3.1. Test de perception et analyse phonologique

Auditeurs guadeloupéens :Les auditeurs guadeloupéens ont commis 105 erreurs sur le /r/ du locuteur guadeloupéen (sur 330 erreurs possibles). Sur ces 105 erreurs, ils ont commis 58 suppressions (soit 55% des erreurs), 21 substitutions (20%), 19 ajouts (18%) et 7 déplacements (7%). A l'écoute du locuteur français, ils n'ont commis que 54 erreurs dont 26 suppressions (48% des erreurs), 21 substitutions (39%) et 7 ajouts (13%) et aucun déplacement.

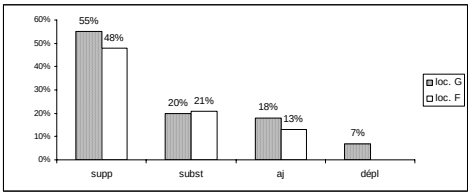


Figure 1 : *Pourcentages des quatre types d'erreurs commis (supp=suppression ; aj=ajout ; subst= substitution) par les 10 auditeurs guadeloupéens sur le /r/ du locuteur guadeloupéen (loc. G, colonnes grisées) et du locuteur français (loc. F)*

Conclusion 1: Les suppressions constituent les principales erreurs commises par les guadeloupéens tant sur le /r/ du locuteur guadeloupéen que sur celui du locuteur français.

Sur 160 réponses sur le /r/ final (16 /r/ en finale*10 auditeurs), les auditeurs guadeloupéens ont commis 50 erreurs (31%) à l'écoute du guadeloupéen et 22 (14%) à l'écoute du français. Sur 80 réponses (8 structures de types Cw et Crw*10 auditeurs) données lors de l'identification des structures de types Cw et Crw, les auditeurs guadeloupéens ont commis 21 erreurs (26%) à l'écoute du guadeloupéen et 5 (6%) à l'écoute du français.

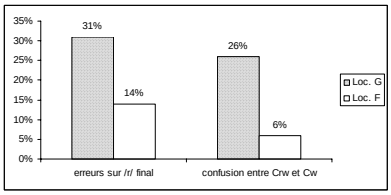


Figure 2 : *Pourcentages des erreurs sur le /r/ final et des confusions faites entre Crw et Cw par les auditeurs guadeloupéens à l'écoute du locuteur guadeloupéen (loc. G, colonnes grisées) et du locuteur français (loc. F)*

Conclusion 2: Les auditeurs guadeloupéens suppriment davantage le /r/ final du locuteur guadeloupéen que celui du locuteur français. Et ils commettent aussi plus de confusions entre les structures de types Crw et Cw à l'écoute du locuteur guadeloupéen que du locuteur français.

Auditeurs français : Les auditeurs français ont commis 140 erreurs sur le /r/ du locuteur guadeloupéen. Sur ces 140 erreurs, ils ont commis 70 suppressions (soit 50% des erreurs), 30 substitutions (22% des erreurs), 27 ajouts (19% des erreurs) et 13 déplacements (9 % des erreurs). A l'écoute du locuteur français, ils n'ont commis que 43 erreurs dont 13 suppressions (30% des erreurs), 23 substitutions (53%), 5 ajouts (12%) et 2 déplacements (5%).

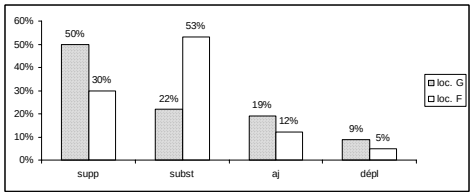


Figure 3 : *Pourcentages des quatre types d'erreurs commis (supp=suppression ; aj=ajout ; subst= substitution) par les 10 auditeurs français sur le /r/ du locuteur guadeloupéen (loc. G, colonnes grisées) et du locuteur français (loc. F)*

Conclusion 3: Tandis que les suppressions constituent le type d'erreurs le plus commis par les français à l'écoute du /r/ des guadeloupéens, ce sont les substitutions qui sont de loin les plus répandues à l'écoute du locuteur français.

Sur 160 réponses sur le /r/ final (16 /r/ en finale*10 auditeurs), les auditeurs français ont commis 64 erreurs (40%) à l'écoute du guadeloupéen et 10 (6%) à l'écoute du français. Sur 80 réponses (8 structures de types Cw et Crw*10 auditeurs) données lors de l'identification des structures de types Cw et Crw, les auditeurs français ont commis 41 erreurs (51%) à l'écoute du guadeloupéen et aucune à l'écoute du français.

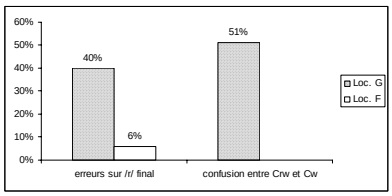


Figure 4 : *Pourcentages des erreurs sur le /r/ final et des confusions faites entre Crw et Cw par les auditeurs français à l'écoute du locuteur guadeloupéen (loc. G, colonnes grisées) et du locuteur français (loc. F)*

Conclusion : Les auditeurs français n'ont aucun mal à distinguer le /r/ uvulaire du français de la semi-consonne /w/. Par contre, ils confondent souvent le /r/ vélaire du guadeloupéen avec le /w/. De même, ils ont du mal à distinguer le /r/ final du guadeloupéen.

3.3. Analyse statistique

Sur la base des données de Troubetzkoy (1939) qui attestent qu'un individu perçoit mieux les sons de son environnement linguistique que ceux d'un milieu étranger, nous émettons les hypothèses H1 et H2.

H1 : les auditeurs guadeloupéens percevraient mieux le /r/ du locuteur guadeloupéen que celui du locuteur français.
H2 : les auditeurs français percevraient mieux le /r/ du locuteur français que celui du locuteur guadeloupéen.

Auditeurs guadeloupéens : Le nombre moyen d'erreurs des auditeurs guadeloupéens sur le /r/ du locuteur guadeloupéen est de 10,5 (avec un écart-type de 1,84) et pour le locuteur français est de 5,4 (avec un écart-type de 2,8). Les résultats du test-t apparié montrent que la différence entre ces moyennes est significative (t= 3,97 et p=0,0033).

Conclusion 4: H1 est infirmée : les auditeurs guadeloupéens ne perçoivent pas mieux le /r/ du locuteur guadeloupéen que celui du locuteur français. Au contraire, les résultats obtenus montrent que les auditeurs créolophones perçoivent beaucoup mieux le /r/ du locuteur français que celui du locuteur guadeloupéen. Des études supplémentaires sont nécessaires, avec plus de locuteurs, car les résultats pourraient être également expliqués par une prononciation plus soignée du locuteur français.

Auditeurs français : Le nombre moyen d'erreurs des auditeurs français sur le /r/ du locuteur guadeloupéen est de 14 (avec un écart-type de 2,2) et pour le locuteur français est de 4,3 (avec un écart-type de 2,7). Les résultats du test-t apparié montrent que la différence entre ces moyennes est significative (t= 13 et p< 0,0001)

Conclusion 5: H2 est confirmée : les auditeurs français perçoivent mieux le /r/ du locuteur français que celui du locuteur guadeloupéen.

L'infirmation de H1 et la confirmation de H2 engendrent de nouvelles interrogations. Sur la base des données de Kuhl et al. (1992) selon lesquelles il s'opère une réorganisation de l'espace perceptif autour des phonèmes de la langue maternelle, nous supposons que le /r/ des français serait plus prototypique pour les auditeurs français que pour les auditeurs guadeloupéens. Nous émettons alors l'hypothèse H3.

H3 : Les auditeurs guadeloupéens percevraient moins bien que les auditeurs français le /r/ du locuteur français.

Auditeurs guadeloupéens et auditeurs français : Le nombre moyen d'erreurs des auditeurs guadeloupéens sur le /r/ du locuteur français est de 5,4 (avec un écart-type de 2,8) et des auditeurs français est de 4,3 (avec un écart-type de 2,7). Les résultats du test-t non-apparié montrent que la différence entre ces moyennes n'est pas significative (t= 0,89 et p= 0,3835).

Conclusion 6: H3 est infirmée : les auditeurs guadeloupéens perçoivent aussi bien que les auditeurs français le /r/ du locuteur français.

4. DISCUSSION ET CONCLUSION

La langue maternelle d'un individu formate tant sa production que sa perception et le dote d'un filtre phonologique à travers lequel il traite la parole. Ce filtre s'accompagne inévitablement d'une surdité phonologique chez l'individu qui, dès lors, n'analyse les sons de la parole (mêmes étrangers à sa propre langue) qu'en se basant sur le système phonologique de sa langue maternelle, ce qui engendre de nombreuses « erreurs » [9].

Les résultats de la présente étude confirment combien la langue maternelle d'un individu a d'influence sur sa perception. Aussi, constatons-nous que bien que ce locuteur bilingue possède une langue (le français) dans laquelle le /r/ est admis dans toutes les positions du mot, cela ne « comble » pas pour autant, dans sa perception, l'absence du /r/ en final en créole.

Il faut aussi retenir que les français se sont montrés beaucoup plus aptes à discriminer le /r/ du locuteur français que celui du locuteur guadeloupéen. Mais, contre toute attente, les guadeloupéens également se sont révélés être plus performants quant à la discrimination du /r/ du locuteur français que de celui du /r/ du locuteur guadeloupéen. Et il semble même que les guadeloupéens (bilingues créole-français) percevraient aussi bien que les français (monolingues français) le /r/ du locuteur français. Cela vient confirmer la double compétence linguistique des bilingues dont a parlé Grosjean [4]. Mais, cela donne aussi à réfléchir sur l'application de la théorie du prototype [6]. Théorie selon laquelle, les langues possèdent des catégories phonétiques qui leurs sont propres et auxquelles est associé un prototype dont la caractéristique est d'être le meilleur exemplaire de la catégorie. Sachant que de part « l'effet magnétique perceptuel » [5] les membres non prototypiques d'une même catégorie phonétique sont perçus comme étant plus proches qu'ils ne le sont en réalité dans l'espace acoustique. Il pouvait alors être supposé que le /r/ uvulaire en tant que membre non prototypique (pour les guadeloupéens) soit moins bien perçu que quand il constitue le prototype de la catégorie phonétique (pour les français). Mais cela ne semble pas être le cas. En revanche, les nombreuses confusions entre le /r/ vélaire guadeloupéen (surtout en contexte labial ; se réalisant /r/ vélaire labialisé) et le /w/ labiovélaire tant pour les auditeurs guadeloupéens que pour les auditeurs français pourraient s'expliquer par le fait que ces deux sons appartiennent à la même catégorie phonétique tant pour les guadeloupéens que pour les français. Cependant, le /r/ vélaire serait plus éloigné du /w/ dans l'espace acoustique pour les guadeloupéens que pour les français. Pour les français, le /r/ vélaire (surtout en contexte labial) serait un son non prototypique d'une catégorie phonétique de laquelle /w/ constituerait le prototype ; ce qui rendrait d'autant plus difficile pour eux l'identification du /r/ guadeloupéen. Tandis que pour les guadeloupéens, le /r/ uvulaire doit se situer dans une catégorie phonétique différente de celle dans laquelle se situe /w/ ; d'où la meilleure discrimination entre ces deux sons.

Enfin, nous posons le problème suivant qui apparaît émerger de nos résultats : comment expliquer que les guadeloupéens aient mieux perçu le /r/ français que le /r/ guadeloupéen bien qu'ils ne sachent pas le produire si ce n'est au prix d'efforts considérables quand ils ne sont pas vains ? Le lien entre production et perception apparaît être bien plus complexe que ce qu'en a pu expliquer la théorie motrice [7] par le fait que l'on serait capable de percevoir ce que l'on peut soi même produire. Ainsi, l'exemple des guadeloupéens nous fait nous demander si le prototype est le meilleur exemplaire du son que l'on perçoit mieux même sans être capable de le produire (/r/ français) ou de celui que l'on produit mieux mais que l'on perçoit moins (/r/ guadeloupéen). C'est ce à quoi nous tenterons de répondre dans une étude ultérieure en travaillant sur un corpus plus vaste en nombre de participants mais aussi en variant les exemples de langues en contact afin de tenter de dégager les tendances universelles quant à l'influence des caractéristiques distributionnelles et acoustiques d'un son sur la perception de la parole chez les monolingues et chez les bilingues.

BIBLIOGRAPHIE

[1] R. Chaudenson. Les Créoles. *Que sais-je?* PUF, Paris, 1995.

[2] A. Cutler, J. Mehler, D. Norris and J. Seguy. The Monolingual Nature of Speech Segmentation by Bilinguals. *Cognitive Psychology* 24: 381-410, 1992.

[3] C. Deprez. *Les Enfants Bilingues: Langues et Familles*. Didier, Paris, 1999.

[4] F. Grosjean. Le Bilinguisme et le Biculturalisme: Essais de Définition. In *Travaux de Neuchâtelois de Linguistique*, Neuchâtel, volume 19, pages 13-35, 1993.

[5] F.H. Guenther and M.N. Gjaja. The perceptual magnet effect as an emergent property of neural map formation. *J. Acoustical Society America*, 100: 1111-1121, 1996.

[6] P.K. Kuhl. Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science*, 255: 606-608, 1992.

[7] A.M. Liberman. Perception of the Speech Code. *Psychological Review*, 74:431-461, 1967.

[8] D. Perani & al. Brain Processing of Native Foreign Languages. *Neuroreport*, 7: 2439-44, 1996.

[9] N.S. Troubetzkoy. *Principes de Phonologie*. Klincksieck, Paris, 1939.

[10] A. Valdman. *Le Créole : Structure, Statut et Origine*. Klincksieck, Paris, 1978.

Vous avez dit *proéminence* ?

Michel MOREL[(1)], Anne LACHERET-DUJOUR[(1)(2)], Chantal LYCHE[(3)], François POIRÉ[(4)]

[(1)] CRISCO, Université de Caen, esplanade de la Paix, 14000 CAEN
[(2)] Institut Universitaire de France, PARIS
[(3)] Université d'Oslo, NORVÈGE
[(4)] The University of Western Ontario, CANADA
morel@crisco.unicaen.fr, lacheret@crisco.unicaen.fr, chantal@ilos.uio.no, fpoire@uwo.ca

ABSTRACT

Analysing prosody requires the correct identification of prominence peaks. This paper examines the results of an experiment where 7 linguists submitted to a prominence identification task largely failed to agree. We show that prominence detection is proportionate to F0 variation, but not to length, that there exists considerable variation between judges, the best of whom barely attains a 50 % score of correct answers. We conclude that coding prosody in a large corpus will require the use of dedicated software to supplement the work done by individual coders.

1. INTRODUCTION

Le travail ici décrit, centré sur l'analyse raisonnée d'une tâche de perception de proéminences par 7 juges phonéticiens dans un extrait de conversation libre, s'inscrit au sein du projet *Phonologie du Français contemporain* (PFC) amorcé en 2002 dont les objectifs initiaux sont présentés dans Durand & Lyche [2]. Dans ce cadre, s'est mis en place un volet prosodique dont une des composantes concerne l'étude de l'interaction des niveaux segmental et suprasegmental (en particulier schwa-prosodie). Cette dernière a progressivement donné lieu à un protocole de codage stabilisé autour de deux modes de codage : standard vs. étendu (Lacheret, Lyche & Morel [4][5]). Comme tout type de transcription prosodique, ce protocole repose en premier lieu sur l'identification de proéminences syllabiques sur des bases auditives. Mais qu'entend-on exactement par ce terme ? Peut-il émerger un consensus lorsqu'on demande à des analystes d'horizons divers d'indiquer les syllabes proéminentes dans un extrait sonore, *i.e.* de coder les événements prosodiques perceptivement remarquables ? La notion même de *syllabe* n'est-elle pas trop restrictive lorsqu'on sait que la proéminence syllabique peut se propager sur les syllabes environnantes (Astesano & al. [1]) ? Enfin, que peut-on dire des points d'ancrage psycho-acoustiques et linguistiques qui sous-tendent la démarche ? En d'autres termes, quelle tâche effectuent réellement les codeurs ? Malgré les meilleures intentions du monde, ces derniers peuvent-ils faire totalement abstraction du niveau symbolique qui conditionne en partie leurs attentes et sous-tend l'émergence des proéminences (structure syntaxique, contraintes sémantiques et informationnelles) ?

En pratique, cette communication fait suite à l'enquête préliminaire de F. Poiré [8], initiateur et fédérateur du travail (mise en service du corpus, définition de la tâche, dépouillement des données). Après avoir rappelé la tâche, sa mise en oeuvre et les premiers résultats dont nous rend compte l'auteur, notre objectif est de mettre en corrélation ces observations avec les configurations acoustiques des données (F0 et durée). Ceci nous permettra de : i) statuer sur la robustesse relative de nos deux paramètres acoustiques pour le repérage des proéminences, ii) proposer des mesures pour évaluer la performance des juges et iii) déterminer le profil des juges qui peut être établi grâce aux deux notions complémentaires de *seuil* et *performance*.

2. LA TACHE ET LES PREMIERES OBSERVATIONS ASSOCIEES

L'objet de cette section est de rappeler les modalités de la tâche envisagée et les premières observations perceptives auxquelles elle a donné lieu.

2.1. La tâche

Un extrait d'environ 3 minutes de parole spontanée produite par un locuteur belge a été choisi pour réaliser la tâche d'identification des proéminences, en position finale et non finale (syllabe proéminente codée '1', syllabe non proéminente codée '0'). Cette tâche a été demandée à 7 sujets, tous phonéticiens et/ou phonologues chevronnés, spécialistes de prosodie. 165 syllabes au total ont été analysées par F. Poiré, soit parce que elles ont été codées '1' par au moins un des 7 juges (ce qui inclut les syllabes emphatiques), soit parce qu'elles sont en position accentuable (syllabes terminales de mots pleins [3]).

2.2. Premiers résultats

Les premières observations formulées par Poiré [8] à l'issue de son analyse s'articulent autour de 3 points majeurs. (i) La perception des proéminences non terminales ou portées par des clitiques reste marginale et pour cause : les sites en question ne correspondent pas aux positions attendues des syllabes accentuables. (ii) Pour l'ensemble des codeurs, la variation dans le pourcentage des syllabes

reconnues comme proéminentes (code '1') est telle (19 % à 49 %) qu'il semble raisonnable de conclure que les sujets ne partagent pas la même définition du concept. (iii) En revanche, l'accord inter-juges est plus sensible pour les syllabes non proéminentes.

3. ANALYSE ACOUSTIQUE DES RESULTATS ET DISCUSSION

3.1. Outils d'analyse

Un relevé de mesure (F0, durée) a été effectué sous PRAAT sur les 165 syllabes analysées ainsi que les syllabes environnantes, utilisées comme référence. La figure 1 présente le traitement d'un extrait de la séquence ci-dessous :

(...) 750 kilos là il te faut un permis B (...)

transcrite en sampa, où l'analyse porte sur la syllabe */lo/* .

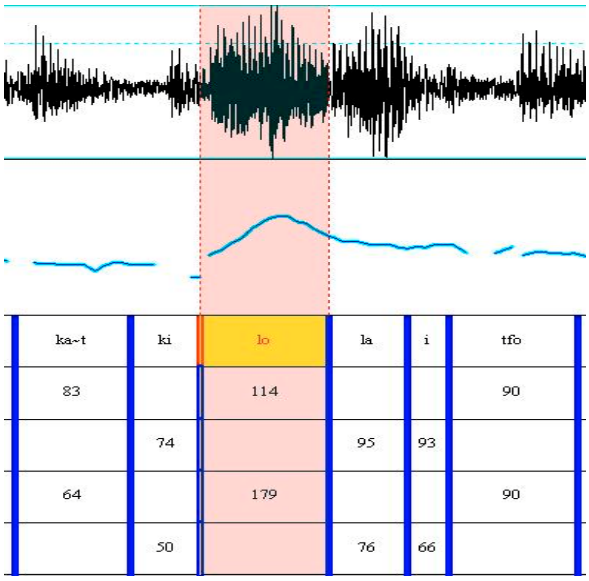


Figure 1 : mesure des proéminences sous PRAAT, illustration pour la syllabe */lo/*. Abscisse : temps. Fenêtre du haut : visualisation du signal sonore. En dessous : courbe F0 correspondante. Fenêtre du bas : 5 tires de traitement remplies manuellement :

- tire 1 : codage phonétique
- tire 2 : 114 Hz = valeur maximale de F0
- tire 3 : 74 et 95 Hz = valeurs environnantes prises comme références
- tire 4 : 179 ms = valeur maximale de durée
- tire 5 : 50 et 76 ms = valeurs environnantes prises comme références.

Sur ces bases, pour chacune de nos 165 syllabes, deux valeurs, *F0* et *durée* (en italique), ont été calculées relativement aux valeurs de référence. *F0* est exprimée en demi-tons (12 unités = rapport 2) selon la formule :

$$F0 = \frac{Log(\frac{F0max}{F0base})}{Log(2)} \times 12$$

F0max étant la valeur maximale de F0 dans la syllabe et *F0base* la moyenne des F0 des syllabes précédente et suivante. Le procédé est le même pour la durée, à ceci près que *durée* est exprimée en pourcentage de la valeur de base. La valeur 100 correspond ainsi à une durée syllabique égale à la moyenne de celle qui précède et de celle qui suit, donc à un allongement nul. Dans notre exemple (syllabe /lo/ de *kilos*), les valeurs calculées sont les suivantes :

F0 = 5,2 demi-tons *durée* = 284 %

3.2. Exploitation des données

Nous avons créé la variable *accord*, qui est la somme des valeurs 0 ou 1 attribuées par les juges. Ainsi, <*accord* = 0> correspond à une syllabe non perçue comme proéminente par l'ensemble des juges alors qu'elle est considérée comme accentuable et <*accord* = 7> à une syllabe perçue proéminente par les 7 juges. Nous avons ensuite mis en correspondance pour chaque syllabe la valeur de *accord* avec les valeurs calculées *F0* et *durée*, afin de comparer perception et mesure (figures 2 et 3).

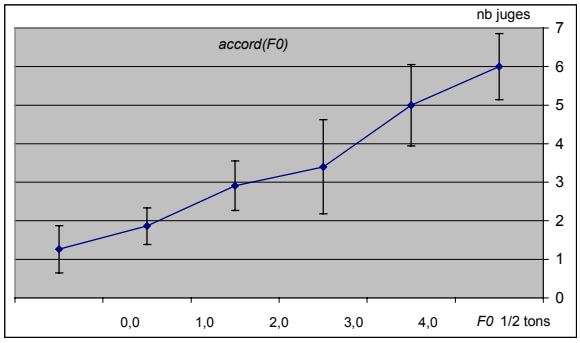


Figure 2 : *F0* en abscisse, *accord* en ordonnée. Chaque point constitue la moyenne des valeurs de *accord* dans la plage considérée de *F0*. Les barres d'erreur représentent les intervalles de confiance à 5 %.

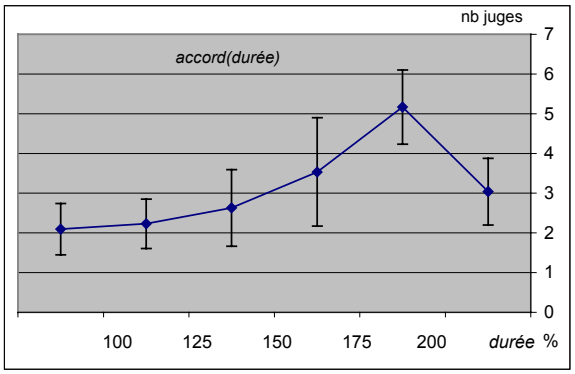


Figure 3 : *durée* en abscisse, *accord* en ordonnée. Chaque point constitue la moyenne des valeurs de *accord* dans la plage considérée de *durée.* Les barres d'erreur représentent les intervalles de confiance à 5 %.

Première remarque : si l'accord des juges est significativement proportionnel aux variations de F0, il n'en va pas de même pour la durée. La figure 2 montre que l'accord des juges est bien corrélé à *F0* dans les valeurs fortes, un peu moins dans les valeurs faibles où il reste de l'ordre de 1 à 2, mais avec toujours le même sens de progression. On peut donc conclure, toute chose égale par ailleurs, que *F0* constitue un corrélat acoustique fiable de la perception des proéminences. La figure 3 en revanche met en évidence une faible corrélation entre *accord* et *durée*, avec une dispersion importante (grands intervalles de confiance). Seules les valeurs de *durée* de 170 à 200 % semblent convaincre plus de la moitié des juges. Mais surtout, les extrémités de ce graphique montrent deux phénomènes qui limitent encore la fiabilité de la durée comme corrélat acoustique de la perception des proéminences. D'une part, des syllabes de faible durée peuvent être perçues comme proéminentes (variation de F0 marquée, allongement nul). D'autre part, les augmentations fortes de durée (supérieures à 200 %) sont souvent associées à des hésitations et de fait perçues comme telles et étiquetées '0'.

Ce dernier point renvoie au concept même de proéminence : si une proéminence est une figure qui se détache sur un fond, alors une durée longue devrait être considérée comme telle. Or la plupart des juges ne considèrent pas l'hésitation comme une proéminence. On peut émettre l'hypothèse qu'ils considèrent – consciemment ou non – sa fonction de gestion du tour de parole comme non pertinente dans le cadre d'une étude sur la prosodie. Quoi qu'il en soit, la question du rôle de la durée dans la notion de proéminence pourrait faire l'objet d'une étude spécifique. Ici, les juges ont appliqué leur propre conception de la proéminence, avec comme résultat une faible corrélation entre durées et proéminence perçues.

La première partie de cette étude confirme la légitimité de sélectionner la F0 comme corrélat acoustique robuste de la proéminence (invariant à la perception de l'ensemble des sujets). Reste à statuer sur la performance individuelle de chaque juge. Autrement dit, qu'en est-il de la variation inter-juges dans la tâche de détection des proéminences ?

Pour chaque juge, nous avons constitué deux sous-ensembles de syllabes : codées '0' vs. '1'. Nous avons rangé les *F0* correspondantes par ordre croissant dans chaque sous-ensemble. La figure 4 présente les résultats obtenus par le juge 3 (courbes croissantes noires) choisi au hasard pour illustrer notre propos. Chez ce dernier, 108 syllabes sont codées '0' et 57 '1'. La référence (courbe croissante grise) correspond à toutes les valeurs de *F0* rangées par ordre croissant. Elle représente donc le résultat virtuel d'un juge *parfait*, qui aurait classé non proéminentes les 108 syllabes où *F0* est la plus basse et proéminentes les 57 syllabes où *F0* est la plus haute. Nous avons appelé *seuil* la valeur de *F0* correspondant à la frontière entre ces deux groupes de syllabes (1,4 demi-tons dans l'exemple choisi figure 4). Bien évidemment, chaque juge possède son propre seuil de perception des proéminences. Celui-ci varie même dans d'assez fortes proportions d'un juge à l'autre. Mais attention : ce seuil ne préjuge en rien de la performance de chacun. (*infra*, table 1, figure 5). En pratique, l'évaluation des performances d'un juge consiste à vérifier s'il a bien perçu comme proéminentes les syllabes dont *F0* est supérieure à son seuil personnel – 1,4 demi-tons pour le juge 3, par exemple – et comme non proéminentes celles dont *F0* est inférieure à ce seuil, autrement dit à tester la cohérence du juge dans la tâche réalisée. Cela revient donc à comparer les résultats de chaque juge à ceux d'un juge *parfait* travaillant avec un seuil de perception identique. Ainsi, pour notre juge 3, la courbe (noire) présente une forte discontinuité à la frontière des syllabes codées '0' et '1'. Ce qui signifie que des valeurs de *F0* hautes n'ont pas été associées à des proéminences, alors que des valeurs basses l'ont été. Quant aux 6 autres juges, ils présentent des caractéristiques similaires, sans pour autant bien sûr que les différences avec le juge *parfait* concernent les mêmes syllabes.

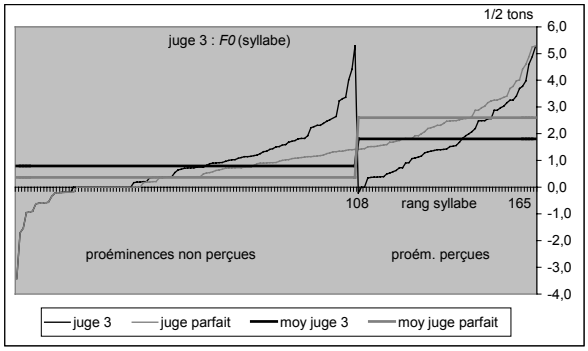


Figure 4 : *F0* en ordonnée, rangée par ordre croissant respectivement sur les syllabes perçues non proéminentes et les syllabes perçues proéminentes.

A ce stade le taux de bonnes réponses ne suffisait pas pour conclure ; il fallait en effet hiérarchiser les erreurs en accordant plus d'importance aux dysfonctionnements manifestes (ex. fortes valeurs de *F0* non perçues comme indices de proéminences et inversement). Pour ce faire, nous avons calculé, pour nos syllabes étiquetées '0' et '1', les moyennes respectives des F0 de chaque juge et du juge *parfait* (plateaux noirs pour notre juge 3 et plateaux gris pour le juge parfait). La différence entre les deux moyennes est plus forte pour le juge *parfait* que pour le juge 3 (respectivement 2,2 et 1,0 demi-tons). En effet, chaque fois que l'on classe une syllabe dans la mauvaise catégorie, les deux moyennes se rapprochent, et ceci d'autant plus que la *F0* de cette syllabe est éloignée du seuil de perception, donc que l'erreur est importante. Nous avons appelé *séparation* la différence entre ces deux moyennes et l'avons utilisée comme indice de performance individuelle des juges. La valeur maximale (MAX) de *séparation* correspond au juge *parfait* ; sa valeur minimale (MIN) correspond à un juge qui classerait aléatoirement les syllabes comme proéminentes ou non ; auquel cas la moyenne des *F0* serait la même dans les deux catégories – non perçues et perçues –, *séparation* serait alors nulle. Concrètement, pour chacun de nos 7 juges, la valeur de *séparation* est comprise entre MIN et MAX.

Table 1 : seuil de perception des proéminences, indice de séparation et performance de chaque juge.

	seuil (1/2 tons)	séparation (demi-tons)	performance (%)
Juge *parfait*	0,9 à 2,3	2,0 à 2,6	100
Juge 1	1,7	1,36	56
Juge 2	0,9	1,05	51
Juge 3	1,4	1,02	46
Juge 4	2,3	1,16	44
Juge 5	1,2	0,91	43
Juge 6	0,9	0,87	42
Juge 7	1,4	0,91	41

La table 1 montre que (i) les seuils de perception des proéminences sont très variables d'un juge à l'autre, (ii) seuls deux juges dépassent une performance de 50 %, ce qui confirme le caractère empirique et en partie aléatoire du jugement individuel perceptif.

Enfin, une représentation graphique des résultats de la table 1, illustrée par la figure 5, confirme l'indépendance entre seuil de perception et taux de performance.

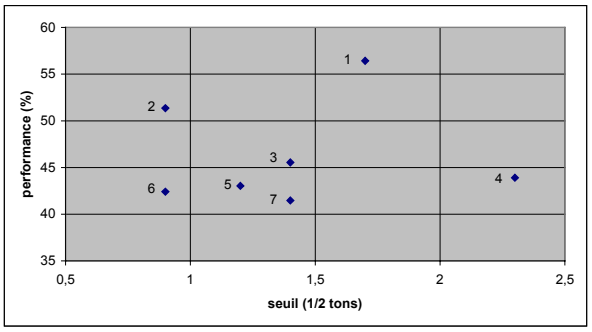


Figure 5 : seuil de détection des proéminences et performance des juges numérotés de 1 à 7.

4. CONCLUSION

Les résultats obtenus à partir de la tâche considérée illustrent plusieurs points. (i) Y compris chez les experts, la notion de proéminence semble loin d'être consensuelle. En conséquence, pour les tâches de codage à venir dans PFC, il paraît nécessaire d'en préciser la portée plus finement, bref, de mieux circonscrire la consigne afin d'aboutir à des étiquetages plus homogènes (par exemple s'entendre sur le traitement des hésitations, également fixer une hiérarchie dans la sélection des proéminences a priori perçues, *i.e.* déterminer un seuil en deçà duquel la syllabe ne peut pas être étiquetée proéminente – voir aussi Martin [6]). (ii) Même dans ces contextes, il est dangereux de se contenter d'un sous-bassement purement auditif pour reconstruire la structure prosodique ; des logiciels d'aide à la détection s'avèrent fondamentalement nécessaires. Au delà de l'utilisation systématique et maintenant classique de l'affichage de F0, voire de la durée, une représentation de la mélodie mesurée, facile à mettre en œuvre, à lire et à interpréter s'avère fort utile (voir par exemple le logiciel Prosogramme développé par Mertens [7]). Une telle approche permettra notamment de contrôler la cohérence des codages obtenus en sachant qu'ils peuvent venir d'horizons fort divers (codeur ± néophyte en phonétique, système dialectal du codeur et du locuteur ± différents), et ainsi de se donner des moyens pour se défaire de la vision pessimiste selon laquelle la détection des proéminences s'apparente plus à un art qu'à une pratique rigoureuse (Martin [6]).

BIBLIOGRAPHIE

[1] C. Astesano, M. Morel, A.L. Coquillon, R. Espesser, M. Besson et A. Lacheret-Dujour. Marquage acoustique du focus contrastif non codé syntaxiquement en français. *25èmes Journées d'Étude sur la Parole*, Fès, Maroc, 19-22 avril 2004.

[2] J. Durand et C. Lyche. Le projet 'Phonologie du Français Contemporain' (PFC) et sa méthodologie. In *Corpus et variation en phonologie du français*. E. Delais-Roussarie et J. Durand (éd.) Toulouse. PUM, pages 213-276, 2003.

[3] A. Lacheret et F. Beaugendre. *La prosodie du français*. Editions du CNRS, Paris, 1999.

[4] A. Lacheret, C. Lyche et M. Morel. Codage prosodique : lien prosodie - schwa/liaison. *Bulletin PFC* n° 3, pages 89-98, 2003.

[5] A. Lacheret, C. Lyche et M. Morel. Pour une transcription prosodique normalisée au sein du projet PFC (Phonologie du français contemporain) : champ d'action et perspectives. *25èmes Journées d'Etude sur la Parole*, Fès, Maroc, 19-22 avril 2004.

[6] P. Martin. La transcription des proéminences accentuelles : mission impossible ? *Bulletin PFC* n° 6, ERSS, UMR 5610, CNRS et Université de Toulouse-Le Mirail, pages 81-87, 2006.

[7] P. Mertens. Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des langues* n° 45 (2), pages 109-130, 2004.

[8] F. Poiré. La perception des proéminences et le codage prosodique. *Bulletin PFC* n° 6, ERSS, UMR 5610, CNRS et Université de Toulouse-Le Mirail, pages 69-79, 2006.

Intonation des phrases interrogatives et affirmatives en langue vietnamienne

Vũ Minh Quang - Trần Đỗ Đạt - Eric Castelli

Centre de recherche international MICA
IP Hanoi – CNRS/UMI-2954 – INP Grenoble
1-Dai Co Viet - Hanoi, Vietnam

{Minh-Quang.Vu ; Do-Dat.Tran ; Eric.Castelli}@mica.edu.vn
http://www.mica.edu.vn

ABSTRACT

Interrogative and affirmative sentences in Vietnamese language, which have same tones and the same number of syllables, are recorded in order to analyse their intonation shape (F0 evolution) avoiding effect of lexical tones and of co-articulation. Comparisons permit us to characterise differences between question and non-question at sentence prosody level. Then, our work is completed by a perception study; its main goal is to check if sentence nature characteristics are included in the sentence prosody, allowing the auditor to classify questions and non-questions, despite the complex form of this prosody in tonal languages. Results show that information on sentence nature are present at the end of its last demi-syllable and that about 70% of sentences are good classified.

1. INTRODUCTION

Pour les langues occidentales non tonales (dont le français ou l'anglais sont des exemples) il a été validé que la prosodie de la phrase véhicule des informations extralinguistiques, comme les émotions ou l'état du locuteur ou bien comme la nature de la phrase (affirmative, interrogative ou exclamative) [1, 2]. Pour évaluer automatiquement le type des phrases à des fins de détection ou de classification, il est alors possible d'analyser directement le signal de parole, sans avoir besoin nécessairement du résultat lexical d'un moteur de reconnaissance automatique de la parole (RAP) mais en utilisant le contour prosodique de la phrase. Dans ce cas, l'essentiel des paramètres mesurés et analysés prennent en compte l'évolution de l'intonation pendant l'énoncé de la phrase : registre de F0, augmentation des valeurs de F0 en fin de phrase ou d'autres paramètres dérivés des valeurs de F0, par exemple [3, 4].

Cependant, dans le cas des langues tonales (comme le mandarin ou le vietnamien), le contour mélodique de l'intonation est complexe parce qu'il est composé de macro-variations correspondant à l'intonation de la phrase et de micro-variations correspondant aux tons lexicaux appliqués sur chacune des syllabes des mots mono (ou bi) syllabiques. C'est pourquoi, appliquer sur ces langues directement les méthodes d'analyse validées pour les langues non tonales semble ne pas permettre de détecter ou de classifier la nature des phrases avec suffisamment de fiabilité car les micro-variations tonales semblent brouiller l'information extra-lexicale de la phrase. Dans le cas de la langue vietnamienne d'ailleurs, pour différencier les phrases interrogatives des autres, l'emploi de mots spécifiques dit « classifieurs interrogatifs » (không, gì, ai, par exemple) est pratiquement systématique. La question alors posée peut être résumée ainsi : existe-t-il, pour le vietnamien, langue à tons dont la prosodie est complexe, des informations extralinguistiques caractérisant le type de phrases, véhiculées par la prosodie et utilisées pendant les actes de dialogue pour la classification de ces types de phrases ? La réponse, outre le fait qu'elle nous permettra d'approfondir nos connaissances de la langue, si elle est positive, nous permettra d'envisager la réalisation de classifieurs automatiques indépendants des moteurs de reconnaissance.

2. ANALYSE DU CONTOUR INTONATIF

2.1. Méthodologie et préparation du corpus

Jusqu'à ce jour, très peu d'études ont analysé la phonologie de la langue vietnamienne en profondeur. Nous pouvons citer quelques travaux récents portant sur les tons lexicaux [5, 6, 7] et sur la prosodie de la phrase [8, 9]. Après analyse de phrases de corpus « lus » et « spontanés », Lê T. X [8] et Nguyễn Thị T. H. & Boulakia [9] constatent qu'il existe une différence de hauteur de F0 entre les types de phrases. En évaluant leur niveau, [8 & 9] précisent que les assertives sont prononcées avec un registre bas alors que les questions et les injonctives le sont avec un registre haut. De plus, [9] fait le constat qu'au niveau de l'allure générale de l'intonation de la phrase, une pente descendante ne correspond pas toujours à une phrase déclarative. Au niveau de la durée : les énoncés interrogatifs ont un débit plus rapide que les énoncés assertifs et injonctifs, la différence de durée entre ces deux derniers n'étant pas significative [9]. Quant à l'intensité, elle est d'une manière générale plus forte dans la phrase interrogative, et les syllabes finales ont souvent un niveau intensité plus important que les autres syllabes de la phrase [9].

Partant de ces constats, nous souhaitons préciser les différences prosodiques entre les phrases interrogatives et

les phrases affirmatives. Pour cela, nous avons construit un corpus spécifique constitué de paires de phrases interrogatives et affirmatives qui présentent, pour chacune des phrases de la même paire, un même contexte tonal et le même nombre de syllabes. Le fait de choisir les mêmes tons nous permet d'éliminer l'influence des tons des syllabes sur l'intonation générale de la phrase, et donc de maîtriser ainsi les micro-variations de l'intonation. Pour éliminer aussi tous les phénomènes de co-articulation qui pourraient interférer avec notre analyse prosodique, nous avons, dans la mesure du possible, gardé aussi les mêmes mots, ou bien nous avons utilisé des mots à la prononciation peu différente. Toutes ces phrases ont été intégrées dans des dialogues significatifs, afin que leur prononciation soit la plus naturelle possible (nous avons enregistré la totalité des dialogues, puis extrait les phrases choisies pour l'analyse). Chaque dialogue est répété cinq fois par six locuteurs (3 hommes et 3 femmes) originaires du Nord du Vietnam. Les phrases sélectionnées sont présentées dans le tableau 1 suivant :

Tableau 1 : Les paires de phrases affirmatives (a) et interrogatives (i) du corpus.

1a	Hôm nay là ngày ba mươi *Aujourd'hui nous sommes le trente*
1i	Hôm nay là ngày bao nhiêu ? *Quel jour sommes-nous aujourd'hui ?*
2a	Tên anh ta là Trì *Il s'appelle Trì*
2i	Tên anh ta là gì ? *Il s'appelle comment ?*
3a	Anh ăn cơm không *Tu manges du riz seulement.*
3i	Anh ăn cơm không ? *Tu manges du riz ?*
4a	Em ăn bánh Ché *Je mange du gâteau Ché.*
4i	Em ăn bánh nhé ? *Tu vas manger du gâteau ?*

En vietnamien, pour la construction de phrases interrogatives, en plus de l'utilisation pratiquement systématique de mots « classifieurs interrogatifs », le locuteur peut ajouter en fin de phrase certain mots qui sont normalement facultatifs mais dont l'usage éventuel dépend fortement de l'habitude, de la façon de parler du locuteur, du contexte dans lequel se produit le dialogue, d'une manifestation de respect et/ou de politesse avec l'interlocuteur, etc. Des exemples sont présentés dans le tableau 2.

Tableau 2 : La même phrase interrogative avec des mots terminaux différents.

- Hôm nay là ngày bao nhiêu rồi ? - Hôm nay là ngày bao nhiêu vậy ? - Hôm nay là ngày bao nhiêu thế ? - Hôm nay là ngày bao nhiêu hả ? → *Quel jour sommes-nous aujourd'hui ?*
- Hôm nay là ngày bao nhiêu hả anh ? → *Quel jour sommes-nous aujourd'hui, monsieur ?*
- Hôm nay là ngày bao nhiêu hả chị ? → *Quel jour sommes-nous aujourd'hui, madame ?*

Comme ces mots terminaux facultatifs peuvent porter n'importe lequel de six tons de la langue vietnamienne, alors la portion finale du contour intonatif de la phrase peut être fortement modifiée par le contour du ton du mot terminal. C'est pourquoi, pour chaque phrase interrogative sélectionnée, nous avons choisi d'incorporer au corpus un certain nombre de variations possédant des mots terminaux de différents tons afin d'étudier plusieurs formes de contours intonatifs possibles.

2.2. Résultats d'analyse

Pour chaque enregistrement nous analysons le contour de la fréquence fondamentale F0 avec le logiciel Praat (exemple présenté figure 1).

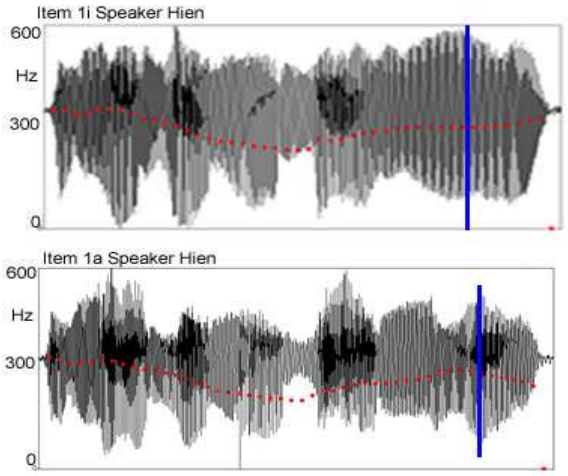


Figure 1 : Deux phrases à nombre de syllabes et tons identiques. En rouge le contour de F0. Figure du dessus : phrase interrogative, dessous : phrase affirmative.

En étudiant chaque paire de phrases présentées dans le tableau 1, nous remarquons que l'essentiel des différences d'intonation se situe à la fin de la phrase (zone située figure 1 après la barre verticale) : le contour de la dernière syllabe ou de la deuxième moitié de celle-ci semble être croissant pour les phrases interrogatives. Une étude statistique, présentée dans le tableau 3, confirme cette tendance : 80% des cas de phrases type I possède un contour de F0 croissant à la fin de phrase. Nous retrouvons là une tendance bien connue pour les langues non tonales comme le français. Cependant, pour un cas de phrase type affirmatif (A), le contour de la dernière moitié de la dernière syllabe est aussi croissant pour 29 enregistrements sur 30. Cette phrase « Em ăn bánh Ché » présente en fin de phrase deux mots au ton croissant qui influence le contour intonatif global de la phrase.

Tableau 3 : Nombre (et pourcentage) des contours de F0 de la dernière moitié de la dernière syllabe.

	Interrogative	Affirmative
Pente montante	96 (80%)	45 (37%)
Pente descendante	24 (20%)	75 (63%)

Nous nous intéressons aussi à la forme intonative des autres phrases de type interrogatives (I) du tableau 2. Bien que certains des mots terminaux portent des tons à contour descendant (ton 2 « descendant » ou ton 6 « grave »), les phrases porteuses de notre corpus présentent en grande majorité (minimum 83 %) un contour croissant à la fin de la phrase (tableau 4).

Tableau 4 : Nombre (et pourcentage) des contours de la dernière moitié de la dernière syllabe pour les phrases interrogatives avec des mots terminaux de différents tons (pas de mots terminaux avec les tons 1 et 3).

	Ton2	Ton6	Ton5	Ton4
Pente montante	45(90%)	80(88%)	75(83%)	28(93%)
Pente descendante	5(10%)	10(12%)	15(17%)	2(7%)

[8 et 9] ont suggéré que les phrases de type I sont prononcées avec un registre plus haut. Pour ce point, l'étude statistique de notre corpus montre qu'effectivement elles semblent montrer une valeur moyenne de F0 plus importante que celle des phrases de type A (tableau 5). Cependant, à part la locutrice Hien qui présente une différence assez importante (environ 40Hz), les différences pour les autres locuteurs sont faibles et plus petites que les valeurs des écarts-type correspondant. En contradiction avec les travaux de [8 et 9], l'effet de registre semble donc peu significatif dans notre corpus. Pour la durée, nous retrouvons la même tendance constatée dans [9] : la durée des phrases interrogatives est en moyenne plus petite de 10% que celle des affirmatives (12% dans [9]).

Tableau 5 : Fréquence fondamentale moyenne (et écart type) des phrases I et A pour les six locuteurs. (H = homme, F = femme).

Locu-teur		Hien F	Quang H	Nam H	Huong F	Diep F	Khoa H
Type I	Hz Hz	307 (16)	160 (15)	160 (8)	250 (18)	239 (11)	145 (16)
Type A	Hz Hz	264 (10)	146 (14)	152 (10)	247 (13)	231 (23)	133 (11)

3. PERCEPTION DES PHRASES I ET A

3.1. Méthodologie et préparation du corpus

Nous souhaitons vérifier que les différences détectées dans notre analyse sont effectivement perçues comme un moyen pour l'auditeur de classifier phrases interrogatives et phrases affirmatives, ou, en d'autres termes, que la prosodie de la phrase, malgré sa complexité due à la présence des tons, véhicule des informations permettant à l'auditeur de faire cette classification. Nous avons utilisé le même corpus décrit ci-dessus auquel nous avons rajouté d'autres phrases affirmatives pour obtenir 13 paires de question/non-question. Pour chaque phrase, après avoir extrait le contour prosodique, nous utilisons ce contour pour synthétiser une pseudo phrase dans laquelle toutes les syllabes sont remplacées par une voyelle unique /a/. Etant donné que la signification sémantique des mots n'existe plus, nous éliminons ainsi la possibilité pour l'auditeur de reconnaître une question uniquement par la présence d'un mot «interrogatif ». Nous reproduisons le plus fidèlement possible, non seulement le contour de l'intonation, mais encore la durée des segments voisés/non voisés et le contour de l'intensité. Puis, nous demandons aux auditeurs de déterminer si la pseudo-phrase synthétisée entendue est interrogative ou affirmative.

3.2. Synthèse des phrases pour le test de perception

L'extraction de ces informations prosodique est réalisée avec le logiciel Prat avec des fenêtres d'analyse de 20 ms pour F0 et 5ms pour l'intensité. Pour la synthèse, nous avons extrait deux périodes de signal de la voyelle /a/ de l'une des phrases prononcées par une locutrice de notre précédente étude.

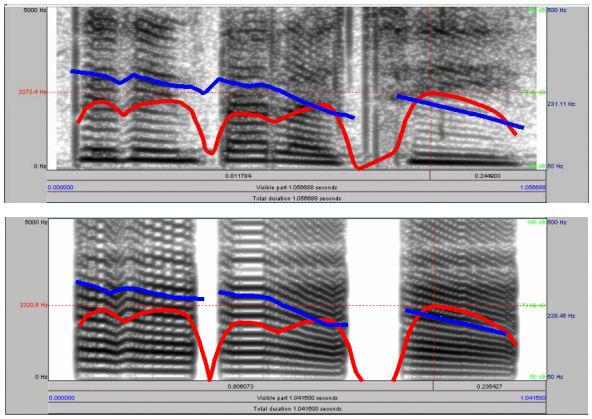


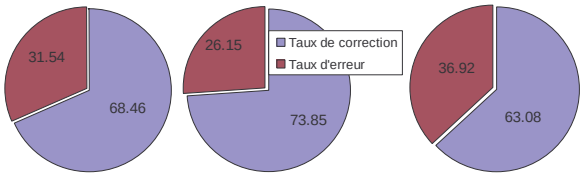
Figure 2 : Spectrogramme, contour de F0 (bleu) et contour d'énergie (rouge) du signal source (a) et du signal synthétisé correspondant (b). Phrase 2a du tableau 1.

L'algorithme TD-PSOLA est utilisé pour concaténer ces extraits de signal, tout en contrôlant le pitch (F0), l'énergie et la durée de chaque syllabe. Dans la phrase synthétisée, les zones correspondant à un signal non voisé (consonnes) sont remplacées par du silence. Nous obtenons ainsi un corpus composé de 13 pseudo-phrases synthétiques « interrogatives » et de 13 pseudo-phrases synthétiques « affirmatives », sans information sémantique. Six auditeurs (3 hommes et 3 femmes) participent à notre test de perception. Ils doivent choisir entre deux réponses « I » ou « A ». Chaque auditeur fait le test 5 fois, et pour chaque session, l'ordre des phrases qui lui sont proposées est aléatoire.

3.3. Résultats de perception

Le résultat du test est présenté dans la figure 2. Le taux de bonne reconnaissance sur l'ensemble des I et A est d'environ 70% (figure 2.a). Les figures 2 .b et 2.c présentent respectivement les taux de bonne classification des phrases interrogatives et des phrases affirmatives : nous pouvons remarquer que les phrases interrogatives semblent mieux reconnues (environ 74 % de bonnes réponses) que les phrases affirmatives (seulement 63%).

La figure 3 détaille les résultats pour les 13 paires de phrase I/A. Pour 10 de ces 13 paires, la phrase interrogative est bien reconnue avec un taux supérieur à 70%), pour la 10[ème] paire, ce taux atteint les 95%. Cependant pour les paires 4 et 12, la phrase affirmative est très mal reconnue (respectivement 12% et 20%).



(a) global (b) interrogatif (c) affirmatif

Figure 3 : Taux de détection correcte : (a) taux global (b) phrases interrogatives et (c) phrases affirmatives.

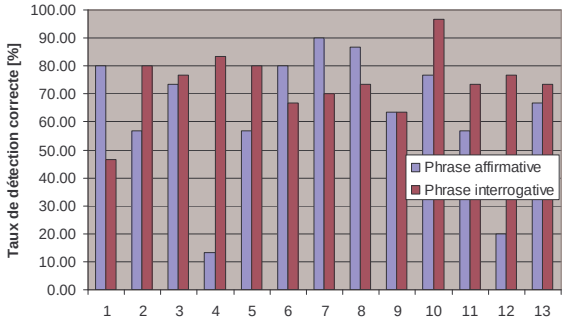


Figure 4 : Taux de détection correcte des 13 paires I/A.

3.4. Discussion

En tentant de corréler ces résultats perceptifs avec ceux de notre analyse de la production des contours intonatifs, nous remarquons que l'auditeur semble juger une phrase comme étant interrogative, si elle présente une intonation croissante en fin de phrase, et juger la phrase comme étant affirmative dans le cas inverse. Cette hypothèse semble être valable pour expliquer le cas des paires 4 et 12 où le taux de reconnaissance des phrases type I est beaucoup plus élevé que celui des phrases type A. Pour ces deux paires, les phrases présentent toutes une dernière syllabe possédant le ton 5 montant, qui fait croître la partie finale du contour intonatif de la phrase, tant pour les interrogations que pour les affirmations. Le fait que le taux de bonne reconnaissance global des phrases A et I soit d'environ 70% (et que pour certaines d'entre elles, elles sont même reconnues à plus de 90%) montre que les paramètres prosodiques de la phrase vietnamienne transportent des informations extralinguistiques qui peuvent permettre à l'auditeur de discriminer le type de phrase. Comme pour les langues non tonales, ces informations sont essentiellement codées par le fait que l'intonation monte ou non en fin de phrase. Cependant, ces informations peuvent être brouillées par la modulation du contour prosodique par les tons lexicaux : des auditeurs peuvent mal classifier des affirmations si les phrases produites présentent une syllabe finale avec ton montant. Des questions peuvent être mal classifiées si leur syllabe finale porte un ton descendant. L'utilisation de mots interrogatifs pour lever les ambiguïtés est donc nécessaire et logique.

4. CONCLUSION

Au niveau production, notre étude a permis de caractériser la prosodie des phrases simples de la langue vietnamienne (dialogue), en éliminant l'influence des tons : les différences entre questions et affirmations sont essentiellement une différence de pente de F0 (croissante ou décroissante) en fin de la phrase (deuxième moitié de la dernière syllabe), à laquelle s'ajoutent une modification du débit. Cependant, pour notre étude, le changement de registre semble plus faible que pour [8 et 9]. Au niveau perceptif, nous avons montré que, comme pour les langues non tonales, la prosodie de la phrase transporte des informations extralinguistiques sur la nature de la phrase, bien que celles-ci, à cause de la présence des tons lexicaux, ne soient pas toujours discriminatives.

BIBLIOGRAPHIE

[1] Rossi M. "*L'intonation, le système du français : description et modélisation*" Editions Ophrys, 1999, ISBN : 2-7080-0912-5

[2] Hirst, D.J. &Di Cristo, A. (Eds.) "*Intonation Systems. A Survey of 20 Languages*" Cambridge: Cambridge University Press.

[3] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M. & Van Ess-Dykema, C. "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech 41*, pp. 439-487, 1998

[4] Vu M.Q., Castelli E., Boucher A. & Besacier L. "Classification de parole en Question et Non-Question par arbre de décision" *SFC 05, 12[èmes] Rencontres de la Société Francophone de Classification* - Montréal, 2005

[5] Nguyen Q.C.**,** Pham Thi N.Y. & Castelli E. "Shape vector characterization of Vietnamese tones and application to automatic recognition" *ASRU 2001 Madonna di Campiglio*, cdrom

[6] Pham Thi N. Y.**,** Castelli E. & Nguyen Q.C. "Gabarits des tons vietnamiens" *JEP 2002 Nancy*, pp 23-26, juin 2002.

[7] Michaud A. & Vu N.T. "Glottalised and non glottalised tones under emphasis: open quotient curves remain stable, F0 curve is modified" *Speech Prosody, Nara, Japan*. 745-748, 2004

[8] Lê Thị X., "Etude contrastive de l'intonation expressive en français et en vietnamien". *Thèse en linguistique : Paris, Université Paris 7*, 1989

[9] Nguyễn Thị T.H. & Boulakia, G. "Another look at vietnamese intonation" *14[th] International Congress of Phonetic Sciences, San Francisco, California*, pp. 2399–2402, 1999

Session VIII

Reconnaissance du locuteur et de la langue

Mardi 13 juin 2006 - 14h30 16h30

Identification automatique des parlers arabes par la prosodie

Jean-Luc Rouas[1], *Melissa Barkat-Defradas*[2], *François Pellegrino*[3], *Rym Hamdi-Sultan*[3]

[1] Laboratoire Électronique Ondes et Signaux pour les Transports (LEOST) - INRETS
[2] Laboratoire ICAR-Praxiling UMR CNRS 5191 - Université Montpellier 3
[3] Laboratoire Dynamique Du Langage UMR CNRS 5696 - Université Lyon 2
jean-luc.rouas@inrets.fr, melissa.barkat@univ-montp3.fr, francois.pellegrino@univ-lyon2.fr, rim.hamdi@univ-lyon2.fr

ABSTRACT

This paper presents a study of automatic identification of Arabic dialectal areas based on a prosodic automatic modelling. Inspired from Fujisaki's works, this modelling dissociates long term prosodic variations from short term micro-variations and exploits n-multigrams models. Experiments, achieved on semi-spontaneous recordings from 40 speakers, show that the system reaches 98% of correct identification of the three dialectal areas - Maghreb, Middle-East, and an intermediate area (Tunisia-Egypt) - with test excerpts of 7.6 seconds in average.

1. Introduction

La prosodie concerne l'ensemble des éléments dynamiques de la chaîne parlée - tels que les variations de hauteur, d'intensité (ou d'énergie) et de durée - qui déterminent la mélodie, les tons, les pauses, les accents, le rythme, le débit...etc. Il s'agit d'un phénomène complexe, relativement difficile à étudier dans la mesure où ses manifestations (i.e. ; patrons accentuels et contours intonatifs) sont sujettes à de nombreuses variations. Les schémas prosodiques diffèrent ainsi selon les langues, les dialectes, les registres linguistiques, la structure du discours, la syntaxe des énoncés, les mots constituant ces énoncés et la structure phonétique des unités lexicales, mais aussi selon le genre, l'âge, l'origine sociale voire l'état émotionnel du locuteur. La complexité de ce phénomène est d'autant plus importante que l'ensemble des ces facteurs interagissent. Une conséquence de cette complexité est que les déterminants de cette variabilité ont été relativement peu étudiés jusqu'à récemment, faute d'outils adaptés. Ainsi, la grande majorité des études cherchaient à neutraliser ces variations pour faire émerger une norme ou en tout cas des patrons moyennés. Or, des études expérimentales récentes montrent que la variation prosodique est un élément de discrimination linguistique et/ou dialectal d'importance. Par exemple, plusieurs études portant sur le rythme ont montré que les formes dialectales d'une même langue pouvaient être discriminées au niveau prosodique sur la base de leur structure rythmique. Des différences inter-dialectales pertinentes ont ainsi été rapportées pour l'anglais [10] et l'arabe [12]. A l'inverse, encore peu d'études ont à ce jour tenté d'évaluer la pertinence des variations intonatives pour identifier des dialectes d'une même langue [11]. Dans cette étude nous nous intéressons aux variations de la fréquence fondamentale (F_0) et de l'énergie dans des énoncés en arabe dialectal afin d'évaluer la pertinence et la robustesse de ces paramètres en identification automatique des parlers arabes par la prosodie.

2. État de l'art des études sur la prosodie en arabe

Alors que l'étude du domaine de l'accent de mot en arabe standard et/ou dialectal est relativement bien développée, les recherches expérimentales sur l'intonation et l'organisation prosodique des énoncés sont relativement peu fréquentes. Les travaux qui abordent l'analyse de la courbe mélodique des phrases concernent le plus souvent l'arabe standard, et ont pour principale application le développement d'outils de synthèse de la parole [15, 23]. La question des patterns prosodiques dans les dialectes arabes modernes a été abordée dans quelques rares études portant plus particulièrement sur les patrons mélodiques associés aux structures syntaxiques et énonciatives les plus courantes (i.e. ; questions totales, assertions) [6, 8, 18]. Les différents parlers arabes n'ont pour autant pas fait l'objet de la même attention. Si les contours intonatifs de l'arabe marocain sont quantitativement bien traités [4, 22], les études s'attachant à la description de ces phénomènes dans les dialectes algériens et/ou tunisiens sont, en revanche, très peu fréquentes [3]. Pour l'heure, nous savons :

- que les parlers arabes présentent des différences significatives au plan des structures syllabiques préférentielles [13], et que la position de l'accent varie en fonction de la structure syllabique [5],
- qu'en arabe dialectal la place de l'accent s'étend de la dernière syllabe à la pré-antépénultième selon le parler considéré [14],
- qu'en arabe marocain au moins, les pics de F_0 qui se concentrent autour des syllabes accentuées connaissent une certaine mobilité du fait de l'influence de certains facteurs (position, type et durée de la syllabe, focus) ce qui correspond aux tendances universelles [22].

Toutefois, compte tenu des résultats prometteurs en identification automatique des parlers arabes par la prosodie présentés dans les sections suivantes, il serait fort utile de s'intéresser à la constitution d'une typologie des contours intonatifs des différents parlers arabe dans une perspective comparative, et ce, afin de mieux cerner la nature des éléments prosodiques de discrimination inter dialectale mis en évidence dans ce travail.

3. Identification automatique des parlers arabes par la prosodie

En s'inspirant du travail d'Adami [1], le système d'identification automatique proposé s'appuie sur un codage des trajectoires de fréquence fondamentale et d'énergie. Cependant, la modélisation se fait ici à plusieurs niveaux, en séparant la fréquence fondamentale en deux composantes : la ligne de base et le résidu, et en proposant des modélisations à des échelles différentes pour ces deux contributions : la ligne de base est modélisée à partir d'unités pseudo-syllabiques, et la modélisation du résidu est fondée sur des unités infra-phonémiques. La fréquence fondamentale et l'énergie sont extraites du signal grâce à la bibliothèque Snack [20]. Le système d'identification automatique est décrit sur la figure 1.

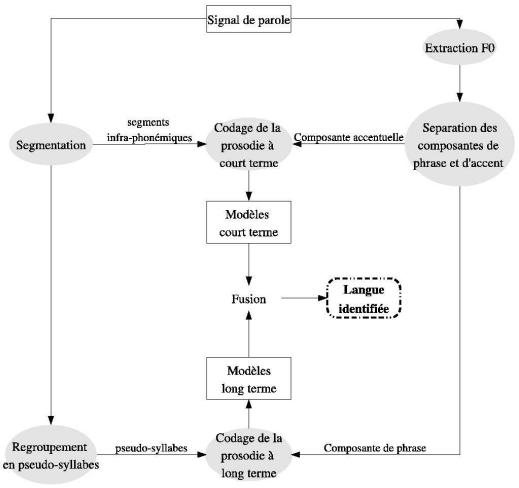


Fig. 1: Description du système

3.1. *Segmentation automatique en unités infra-phonémiques et regroupement en unités pseudo-syllabiques*

Trois traitements de base conduisent à la localisation de frontières quant à la notion de consonne/voyelle :
- segmentation automatique de la parole en segments quasi-stationnaires [2],
- détection d'activité vocale
- localisation des voyelles [16].

Les segments vocaliques sont étiquetés "V", les segments de non-activité "#", et les autres segments "C". Ces segments, de taille infra-phonémiques, seront utilisés pour l'étiquetage des variations locales de fréquence fondamentale et d'énergie.

Ces segments sont regroupés en unités pseudo-syllabiques : la syllabe est en effet une unité privilégiée pour la modélisation du rythme. Néanmoins, la segmentation automatique en syllabes (en particulier en ce qui concerne la détection des frontières) est une opération délicate et spécifique à chaque langue [17]. Pour cette raison, nous utilisons la notion de pseudo-syllabe [19]. Le signal de parole est segmenté en motifs correspondant à la structure [CC...CV]. Un exemple de segmentation automatique en pseudo-syllabes est donné dans la figure 2.

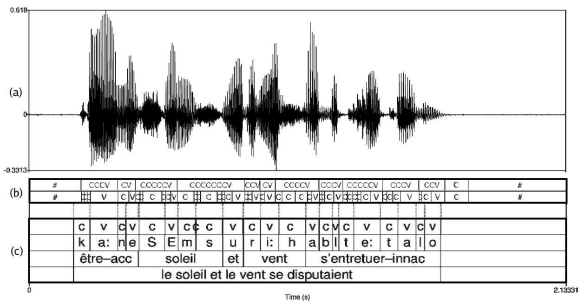


Fig. 2: Segmentation en pseudo-syllabes pour un enregistrement d'un locuteur libanais. (a) signal (b) transcription automatique (c) transcription manuelle

3.2. *Traitement de F_0*

En s'inspirant des travaux de Fujisaki [9], le traitement de la fréquence fondamentale est divisé en deux phases permettant de représenter l'accentuation de phrase et l'accentuation locale.

Accentuation de phrase Il est supposé que la ligne de base passe par les minima locaux de F_0, de telle sorte qu'aucun point ne se situe au-dessous d'elle [21]. Pour chaque phrase, le même traitement est appliqué :
- les valeurs de la fréquence fondamentale en Hertz sont converties en demi-tons. Cette quantification permet de se reporter sur une échelle logarithmique, proche de l'échelle de la perception humaine, et de lisser la courbe mélodique.
- la droite de régression linéaire est estimée sur l'ensemble des parties voisées de chaque phrase. Le minimum est alors repéré sur chaque partie voisée située en dessous de cette droite.
- l'accentuation de phrase ou ligne de base est la droite qui rejoint les minima de la phrase.

La pente de la régression est utilisée pour étiqueter la ligne de base. Une étiquette est employée pour chaque pseudo-syllabe : "U" pour une pente positive et "D" pour une pente négative. Les pseudo-syllabes considérées comme peu ou pas voisées (dont le pourcentage en durée de voisement n'excède pas 70%) sont étiquetées "#".

Un exemple d'extraction de la ligne de base est représenté figure 3. Sur cet exemple, la séquence d'étiquettes correspondant à la phrase est : `D.D.#.D.D.-D.-#.U.#.D.D.#`

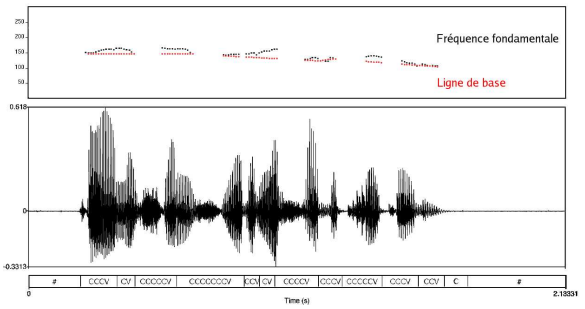


Fig. 3: Extraction de la ligne de base, sur le même exemple

Accentuation locale La ligne de base est soustraite du contour original de F_0. La courbe résultante est appelée résidu (figure 4). Le résidu est approximé sur chaque segment infra-phonémique par une régression linéaire. La pente de l'approximation linéaire est utilisée pour étiqueter les mouvements de F_0 avec trois étiquettes possibles ("U", "D" et "#"). La phrase donnée en exemple donne l'étiquetage suivant : `D.-D.U.D.D.#.#.U.D.D.U.D.#.#.#.U.D.U.U.U.U.D.-#.U.U.D.#.#.D.D.#.#.#.#.U.D.#.#.U.D.U.D.D.#`

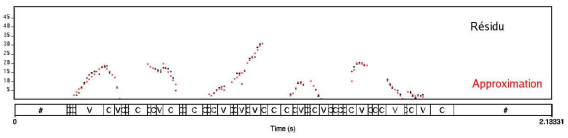


Fig. 4: Approximation du résidu, toujours sur le même exemple. Une approximation linéaire est effectuée sur chaque segment.

3.3. Codage de l'énergie

De la même manière, une régression linéaire est effectuée sur la courbe d'énergie, pour chaque segment infra-phonémique. Pour chacun de ces segments, une étiquette ("U", "D" ou "#") est apposée à la courbe d'énergie.

3.4. Codage de la durée

Les étiquettes de durée ne sont calculées que pour les unités infra-phonémiques. Les étiquettes sont assignées en considérant la durée moyenne selon le type de segment (vocalique, non vocalique ou pause). Par exemple, si un segment vocalique possède une durée supérieure à la durée moyenne des segments vocaliques des données d'apprentissage, il est étiqueté "l" (long). Sinon, il est étiqueté "s". Les étiquettes de durée générées sur la phrase d'exemple sont : `l.s.s.s.l.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.-s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.s.-s.s.l.s.s.l.l.l`

3.5. Modélisation

Les enchaînements des étiquettes sur les phrases de l'ensemble d'apprentissage sont modélisés par des modèles multigrammes. Les séquences les plus fréquentes sont alors identifiées pour chaque langue et des probabilités leur sont associées. Un modèle de langage multigrammes est un modèle statistique qui probabilise chaque motif de suites d'unités [7]. Pour ce système, cette modélisation est utilisée à deux niveaux :

- en employant les étiquettes de variation de F_0 pour la ligne de base, à l'échelle de la pseudo-syllabe. Les modèles multigrammes permettent de travailler sur les enchaînements de plusieurs pseudo-syllabes, ce qui correspondrait à l'échelle du mot.
- en utilisant les étiquettes de variation de F_0 sur le résidu, conjointement avec les étiquettes de durée et d'énergie. Ici, les multigrammes permettent de modéliser les enchaînements de plusieurs segments infra-phonémiques, se rapprochant de l'échelle syllabique.

4. Expériences d'identification automatique

Les données utilisées ont été collectées entre 1995 et 2005 au laboratoire Dynamique Du Langage (corpus Araber). Il s'agit d'enregistrements semi-spontanés (commentaire d'une histoire en images, ...) de 40 locuteurs issus du Maghreb, du Moyen-Orient et de la zone dite intermédiaire (Tunisie, Egypte). Vu la relativement faible quantité de données, tous les enregistrements disponibles pour les locuteurs (soit 5 min. par locuteur, réparties en 40 fichiers de 7,6 secondes de durée en moyenne) ont été utilisés, à la fois en apprentissage et en test, via une procédure de validation croisée (procédure de test d'un seul locuteur en utilisant tous les autres pour apprendre les modèles, itérée pour chaque locuteur, soit 39 fois). Ces expériences sont toutefois à considérer avec précaution, puisqu'elles ne permettent pas d'évaluer la dépendance des modèles vis à vis du texte (les mêmes types de texte apparaissant en apprentissage et en test).

4.1. Résultats de la modélisation à long terme

En ne considérant que les variations de F_0 de la ligne de base, le système permet d'obtenir en moyenne 54 % d'identifications correctes (tableau 1) en utilisant des 5-multigrammes, c'est-à-dire en modélisant les enchaînements jusqu'à 5 pseudo-syllabes consécutives.

Tab. 1: Nombre de fichiers correctement identifiés/testés = 859/1592 (en %)

	Zone occident.	Zone interméd.	Zone orientale.
Zone occident.	49.3	29.0	21.7
Zone interméd.	20.2	59.0	20.8
Zone orientale	18.2	27.1	54.7

Le taux d'identification correcte est significativement supérieur au hasard, mais il est décevant. Les dialectes du groupe « intermédiaire »(i.e., tunisien & égyptien) sont les mieux identifiés mais plus de 40 % des fichiers demeurent mal classés. La courbe de base de l'intonation ainsi modélisée sur une période relativement longue (5 pseudo-syllabes correspondent à une durée d'environ 500 ms) semble donc peu discriminante.

4.2. Résultats de la modélisation à court terme

En considérant les étiquettes de F_0 pour le résidu, l'énergie et la durée des segments, le système permet d'obtenir en moyenne 98 % d'identifications correctes (tableau 2) en utilisant des 5-multigrammes, c'est-à-dire en modélisant les enchaînements jusqu'à 5 segments consécutifs.

Cette modélisation à court terme permet d'obtenir de très bonnes performances quelle que soit la zone dialectale. Le fait que les dialectes de l'est et de l'ouest

Tab. 2: Nombre de fichiers correctement identifiés/testés = 1563/1592 (en %)

	Zone occident.	Zone interméd.	Zone orientale
Zone occident.	99.5	-	0.5
Zone interméd.	1.8	96.0	2.2
Zone orientale.	1.0	0.3	98.7

sont légèrement mieux identifiés que ceux de la zone intermédiaire est dû dans une très large mesure à un unique locuteur, responsable à lui seul de 3/4 des erreurs.

5. Conclusion

Cette étude visait à évaluer la pertinence d'une modélisation prosodique automatique pour l'identification de zones dialectales de l'arabe. Cette modélisation prend en compte des variations de Fo à relativement long terme ainsi que des micro-variations de Fo et d'énergie à court terme. Les expériences, menées avec 40 locuteurs permettent d'atteindre un taux d'identification correcte de 98 % en considérant 3 zones dialectales, à savoir le Maghreb, le Moyen-Orient, ainsi qu'une zone intermédiaire (Tunisie-Egypte). Si les modèles à court terme obtiennent de très bons résultats probablement liés à la prise en compte d'interactions entre accentuation et structure syllabique, les modèles à plus long terme se révèlent peu discriminants. Il est probable que l'échelle temporelle utilisée dans le modèle n-multigramme (supérieur à la syllabe mais inférieur à la proposition) ne soit pas la plus adaptée à des phénomènes intonatifs de plus grand empan temporel.

Références

[1] A. Adami, R. Mihaescu, D.A. Reynolds, and J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *ICASSP*, volume 4, pages 788–791, Hong Kong, China, 2003.

[2] R. André-Obrecht. A new statistical approach for automatic speech segmentation. *IEEE Trans. on ASSP*, 36(1) :29–40, 1988.

[3] I. Benali. Le rôle de la prosodie dans l'identification de deux parlers algériens : l'algérois et l'oranais. In *Workshop MIDL*, pages 128–132, Paris, 29-30 novembre 2004.

[4] T. Benkirane. *Intonation Systems : a Survey of Twenty Languages*, chapter Intonation in Western Arabic (Moroccan). D. Hirst & A. Di Cristo, Eds, Cambridge University Press, 1996.

[5] R. Bouziri, H. Nejmi, and M. Taki. L'accent de l'arabe parlé à Casablanca et à Tunis : étude phonétique et phonologique. In *ICPhS*, pages 134–137, Aix-en-Provence, 1991.

[6] D. Chahal. A preliminary analysis of Lebanese Arabic Intonation. In *Conference of the Australian Linguistic Society*, pages 1–17, 1999.

[7] S. Deligne and F. Bimbot. Language modeling by variable length sequences : theorical formulation and evaluation of multigrams. In *ICASSP*, pages 169–172, Detroit, MI, September 1995.

[8] Sh. El Hassan. The intonation of questions in English and Arabic. *Papers and Studies in Contrastive Linguistics*, pages 97–108, 1998.

[9] H. Fujisaki. Prosody, information and modeling - with emphasis on tonal features of speech. In *ISCA Workshop on Spoken Language Processing*, Mumbai, India, January 2003.

[10] E. Grabe and E.L. Low. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology 7*, 2002.

[11] M. Grice, M. D'Imperio, M. Savino, , and C. Avesani. *Prosodic Typology : The Phonology of Intonation and Phrasing*, chapter A strategy for intonation labelling varieties of Italian, pages 362–389. Oxford : OUP, 2005.

[12] R. Hamdi, M. Barkat-Defradas, and F. Pellegrino. De la caractérisation linguistique à l'identification automatique des dialectes arabes. In *Workshop MIDL*, Paris, 29-30 novembre 2004.

[13] R. Hamdi, S. Ghazali, and M. Barkat-Defradas. Syllable Structure in Spoken Arabic : a comparative investigation. In *Eurospeech*, pages 2465–2468, Lisboa, 2005.

[14] D.E. Khouloughli and G. Bohas. *Analyse et Théories*, volume 1, chapter Processus accentuels en arabe (parlers du Caire, de Damas & arabe classique), pages 1–59. 1981.

[15] M. Mahwoub. Prosodie et ordre des constituants dans l'énoncé en arabe standard moderne. In *JEP-TALN*, Fès, Maroc, 2004.

[16] F. Pellegrino and R. André-Obrecht. Vocalic system modeling : A vq approach. In *IEEE Digital Signal Processing*, pages 427–430, Santorini, July 1997.

[17] H.R. Pfitzinger, S. Burger, and S. Heid. Syllable detection in read and spontaneous speech. In *ICSLP*, volume 2, pages 1261–1264, Philadelphia, October 1996.

[18] J. Rosenhouse. Features of Intonation in Bedouin Arabic narratives of the Galilée (Northern Israel). In *Dialectologia Arabica, a Collection of articles in honor of Professor Heikki Palva*, volume 75, pages 193–215. Studia Oientalia, 1995.

[19] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht. Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, 47(4) :436–456, 2005.

[20] K. Sjölander. The snack sound toolkit. *http ://www.speech.kth.se/snack/*.

[21] J. Vaissière. Language independent prosodic features. In *Prosody : models and measurements*, Springer series in language and communication, 14, pages 53–66. Cutler, A. and Ladd, D.R. (eds.), Berlin, 1983.

[22] M. Yéou. Effects of focus, position and syllable structure on F0 alignment patterns in Arabic. In *JEP*, Fès, Maroc, 2004.

[23] A. Zaki, A. Rajouani, and Z. Najim. Contours intonatifs de la phrase interrogative en arabe. In *JEP*, pages 249–257, Aussois, Suisse, 2000.

Identification automatique des langues: combinaison d'approches phonotactiques à base de treillis de phones et de syllabes

Dong Zhu, Martine Adda-Decker

LIMSI-CNRS, Université de Paris Sud,
BP 133, 91403 Orsay Cedex, France
dong.zhu@limsi.fr, madda@limsi.fr

ABSTRACT

This paper investigates the use of phone and syllable lattices to automatic language identification (LID) based on multilingual phone sets (73, 50 and 35 phones). We focus on efficient combinations of both phonotactic and syllabotactic approaches. The LID structure used to achieve the best performance within this framework is similar to PPRLM (parallel phone recognition followed by language dependent modeling) : several acoustic recognizers based on either multilingual phone or syllabe inventories, followed by language-specific n-gram language models. A seven language broadcast news corpus is used for the development and the test of the LID systems. Our experiments show that the use of the lattice information significantly improves results over all system configurations and all test durations. Multiple system combinations further achieve improvements.

1. INTRODUCTION

L'identification automatique des langues (IAL), qui consiste à déterminer la langue utilisée par un locuteur inconnu est un domaine de recherche très actif. Depuis une dizaine d'années, l'approche PPRLM [1] [2] [5] [9], qui consiste à mettre plusieurs décodeurs acoustiques en parallèle, utilise les modèles acoustiques à base de HMM (Hidden Markov Models), s'avère très performante pour l'IAL. Cette approche exploite de manière implicite un niveau acoustico-phonétique et de manière explicite un niveau phonotactique (via les modèles de langage n-grammes). D'autres niveaux peuvent contribuer à identifier la langue. Des recherches récentes montrent que la combinaison de différents niveaux d'information peut améliorer les performances d'identification. Ainsi les indices prosodiques peuvent contribuer à l'IAL [10], et l'approche PPRLM s'améliore avec l'intégration des modèles HMMs prosodiques [11] ; l'utilisation de SVM (Support Vector Machines) exploite la capacité de classification discriminante du système d'IAL [12] et pour l'évaluation du NIST (National Institut of Standards and Technology, les Etas-Unis) 2003, le meilleur système d'IAL est celui qui combine trois approches : PPRLM, GMM et SVM. Récemment, l'introduction de treillis de phones à la place de la meilleure séquence de phones décodée et l'utilisation de réseaux de neurones dans le module de décision [5] ont donné les meilleurs résultats sur le corpus NIST 2003.
Dans nos expériences récentes [7], nous avons présenté la modélisation de jeux de phones multilingues, l'introduction de l'approche syllabotactique pour l'IAL, ainsi que la supériorité des modèles acoustiques en contexte sur ceux indépendants du contexte. Les travaux antérieurs donnent aussi une comparaison préliminaire entre l'approche phonotactique et syllabotactique.
Les études présentées ici s'inscrivent dans la continuité de nos expériences précédentes. Elles essayent d'estimer l'apport de différentes configurations de système d'IAL et de combinaisons de plusieurs approches. Les travaux se concentrent sur plusieurs questions. D'abord, pour le système d'IAL à modèles acoustiques multilingues, l'utilisation de treillis peut-elle améliorer la représentativité des modèles de langage n-grammes ? Ensuite, concernant l'approche syllabotactique, est-ce que sa combinaison avec l'approche phonotactique peut améliorer les performances d'identification par rapport aux deux autres approches ? Si oui, pour la structure PPRLM, quelle est la meilleure combinaison de décodeurs phonétiques et syllabiques ?

2. CORPUS ET APPROCHE GÉNÉRAL

2.1. Corpus

Un corpus multilingue d'émissions de radio et de télévision a été collecté. Ce corpus offre plusieurs avantages pour l'IAL : la grande quantité de données est favorable à l'apprentissage du système ; la qualité du corpus est intéressante pour l'apprentissage des modèles acoustiques multilingues. Nous utilisons des corpus en sept langues : arabe classique, anglais américain, allemand, espagnol, français, italien, chinois mandarin, avec environ 20 heures par langue. Les corpus français et arabe sont des ressources fournies par la DGA. L'anglais, l'espagnol et le chinois sont extraits des corpus HUB4 du LDC. Les corpus allemands et italiens sont issus de divers projets européens FP5 LE (Olive, Alert) ou obtenus auprès d'ELDA. Pour ces corpus, des transcriptions orthographiques sont disponibles, et des lexiques de prononciation (phonémique, syllabique) ont été adaptés selon le choix de phones multilingues.
Le corpus d'apprentissage est divisé en deux parties. La moitié du corpus, soit environ 10 heures par langue, sert à l'apprentissage des modèles acoustiques multilingues. Le reste du corpus est utilisé pour l'apprentissage des modèles de langage spécifiques à chaque langue. Les corpus de test sont d'environ 30 minutes par langue, et ils sont divisés en segments de différentes durées : 3 secondes, 10 secondes, 20 secondes et 30 secondes.

2.2. Inventaires multilingues phonétiques et syllabiques

Concernant les jeux de phones multilingues, ils ont été définis à partir de huit jeux de phonèmes dépendants de la langue. Ici la huitième langue est le portugais, que nous n'avons pas gardé pour les tests, par manque de corpus adéquats. Des regroupements ont été effectués sur des critères linguistiques pour aboutir à un inventaire de soixante-treize phones multilingues (29 voyelles,40 consonnes). Pour la suite, nous avons exploré différents regroupements et les classifications des phones multilingues pour des études comparatives avec un nombre variable de phones. Ainsi, nous avons regroupé 29 voyelles en 6 classes en fonction de leurs valeurs de formants (F1, F2), cela donne un jeu de 50 phones gardant de nombreuses distinctions de consonnes, mais seulement 6 voyelles. Ensuite, la similarité des lieux d'articulation est la règle utilisée pour diminuer le nombre de consonnes. Finalement un jeu de 35 phones est obtenu qui contient 6 voyelles, 25 consonnes, et 4 phones spéciaux pour modéliser silence, hésitation, bruit, et respiration (table 1).
Le recensement des syllabes est effectué sur le corpus transcrit en syllabe d'après les règles de syllabation établies pendant les travaux précédents [7]. Les dictionnaires syllabiques ont été définis (9788 syllabes pour 73 phones, 9536 syllabes pour 50 phones, et 7712 syllabes pour 35 phones), en fonction de critères d'équilibre de répartition des syllabes sur les huit langues, et de taux de couverture. La fréquence d'occurrence comme critère de sélection permet d'éliminer syllabes rares, difficiles à modéliser avec une approche statistique, ainsi que des syllabes erronées.

2.3. Approche générale

Pour les trois jeux de phones multilingues décrits précédemment, des modèles acoustiques en contexte ont été estimés (3843 modèles acoustiques en contexte pour les 73 phones, 3455 modèles pour les 50 phones, 3415 modèles pour les 35 phones). Les modèles acoustiques sont des modèles HMMs à trois états, et chaque état contient trente-deux gaussiennes. Des modèles de langage phonétiques et syllabiques tri-grammes sont estimés à partir de données issues du décodage sans contrainte de modèles de langage. Les décodeurs acoustiques servent simplement à transformer les signaux acoustiques de parole en une suite de phones ou de syllabes. L'identité de la langue est ensuite obtenue en calculant le maximum de vraisemblance entre les modèles de langage spécifiques à chaque langue et les unités issues du décodage. Au lieu de se limiter à la meilleure solution, l'approche par treillis permet d'exploiter également les hypothèses sous-optimales générées lors du décodage acoustico-phonétique (acoustico-syllabique). L'utilisation de treillis permet ainsi d'améliorer la représentativité des hypothèses produites par les décodeurs acoustiques, en maximisant l'espérance du logarithme de la vraisemblance de la séquence des phones (ou syllabes) décodés (équation 1). L* représente la langue identifié, H représente une des séquences de phones (ou de syllabes) décodées et L représente la langue correspondante.

$$L^* \simeq \underset{L}{argmax}\ E_H[\log P(H|L)] \qquad (1)$$

Deux types de structures d'IAL sont proposés ici pour l'IAL : PRLM (Phone Recognition Followed by Language Modeling) et PPRLM. Dans le cadre du PRLM, neuf systèmes d'IAL ont été mis en place : (1) trois systèmes (73, 50 et 35 phones) d'IAL phonotactiques sans treillis, (2) trois systèmes d'IAL phonotactiques à base de treillis, (3) trois systèmes d'IAL syllabotactiques à base de treillis. Dans le cadre du PPRLM, les systèmes d'IAL bi-décodeurs, tri-décodeurs, quadri-décodeurs (figure 1) et penta-décodeurs sont mis en œvre pour isoler le système d'IAL le plus performant.

TAB. 1: Tableaux des jeux de phones multilingues.

Phones	73	50	35
Voyelles	a ʌ$_{gp}$ ʌ$_{ce}$ a : ã$_f$ ã$_p$	a	a
	e ɛ ɛ̃ æ ə : ɛ : ɜˑ ɣ	e	e
	i ĩ Ɪ ɨ	i	i
	o ɔ õ	o	o
	u ũ ʊ	u	u
	y Y ə ø ʔ	y	y
Semi-voyelles	w	w	w
	ɥ	ɥ	ɥ
	j	j	j
Consonnes	c ç	c ç	c
	ʧ	ʧ	ʧ
	s ʂ θ	s ʂ θ	s
	z ʐ ð	z ʐ ð	z
	ʃ	ʃ	ʃ
	ʒ	ʒ	ʒ
	f	f	f
	v	v	v
	m	m	m
	m̩	m̩	m̩
	n	n	n
	n̩	n̩	n̩
	l l̩	l l̩	l
	l̩	l̩	l̩
	R r χ h ħ r$_e$	R r χ h ħ r$_e$	r
	p	p	p
	b	b	b
	t ʈ	t ʈ	t
	d ɖ	d ɖ	d
	k q	k q	k
	g	g	g
	ɲ	ɲ	ɲ
	ɳ	ɳ	ɳ
spéciaux	.! & W	.! & W	.! & W

3. EXPÉRIENCES MENÉES

Des expériences ont été menées afin de comparer les performances de différentes combinaisons et approches. Au total, 19 systèmes d'IAL sont construits. Les tests se font sur les segments de 3, 10, 20 et 30 secondes, et les données de test distinctes des données d'apprentissage sont issues des mêmes types de sources que ces données. Les corpus du test, soit environ une heure par langue, sont divisés en segments de durées variées, 3 secondes (de 3s à 5s), 10 secondes (de 10s à 12s), 20 secondes (de 20s à 25s), 30 secondes (de 30s à 35s).

3.1. Différents jeux de phones multilingues

Comme le montrent les figure 2 et 3, parmi les systèmes d'IAL de différents jeux de phones multilingues, le sys-

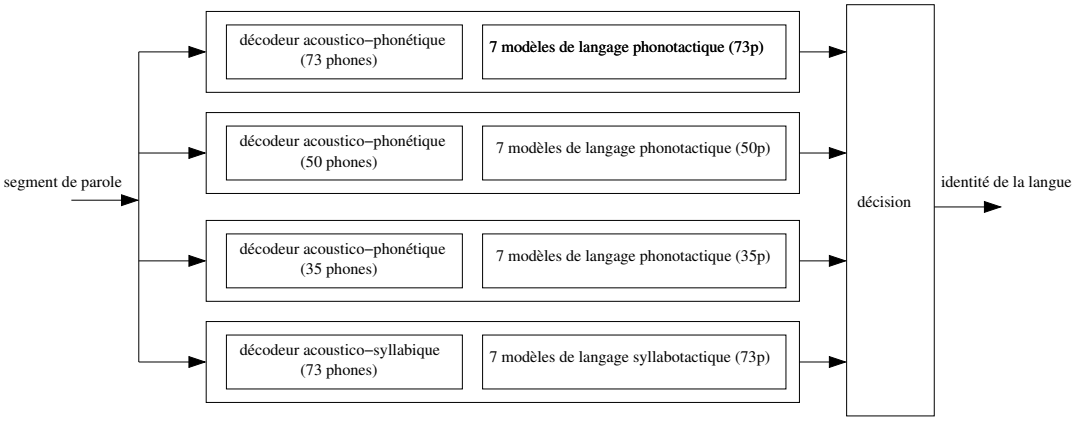


FIG. 1: Le système d'IAL (PPRLM) le plus performant combine l'approche phonotactique et l'approche syllabotactique utilisant les modèles acoustiques multilingues en

tème de 73 phones obtient le meilleur résultat (au niveau phonotactique aussi bien que syllabotactique). Les systèmes d'IAL de 73 et 50 phones sont généralement meilleurs que les systèmes de 35 phones, pour lequel les regroupements consonantiques doivent être remis en question.

3.2. Utilisation de treillis

Comme le montre la figure 2, sur tous les segments de test (3s, 10s, 20s, 30s) les systèmes phonotactiques d'IAL à base de treillis sont meilleurs que les systèmes sans treillis. On constate que sur les segments de test courts (3 secondes), le système d'IAL de 73 phones à base de treillis obtient un gain de 0,9% sur le système d'IAL de 73 phones sans treillis ; des gains peuvent également être mesurés pour le système d'IAL de 50 phones (0,2%) et le système de 35 phones (0.9%). Sur les segments de test longs (30 secondes), le gain est plus fort pour les systèmes avec moins de phones (2,5% pour 35 phones) que pour ceux avec plus de phones (0,3% pour 50 phones, 0,3% pour 73 phones). Les taux d'erreur d'identification sont présentés en détail dans la table 2.

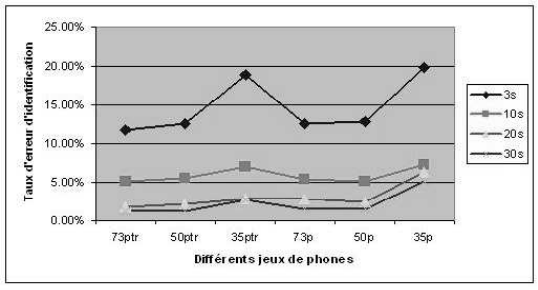


FIG. 2: Taux d'erreur d'identification des systèmes phonotactiques avec ou sans treillis, 73ptr signifie le système phonotactique de 73 phones avec treillis, 73p signifie 73 phones sans treillis.

3.3. Utilisation de l'approche syllabotactique

Le système d'IAL syllabotactique utilise les mêmes modèles acoustiques multilingues que le système phonotactique, mais le décodeur acoustico-syllabique produit des syllabes au lieu des phones, et les modèles de langage syllabotactiques se basent sur des milliers de syllabes. Nous avons illustré dans la figure 3 les résultats d'identification des systèmes syllabotactiques avec treillis et pour comparaison, nous avons ajouté les résultats des systèmes phonotactiques. Au niveau syllabotactique, le système d'IAL de 73 phones est le plus performant sur tous les segments de test (3s, 10, 20s et 30s). Il reste cependant un peu moins performant que le meilleur système phonotactique.

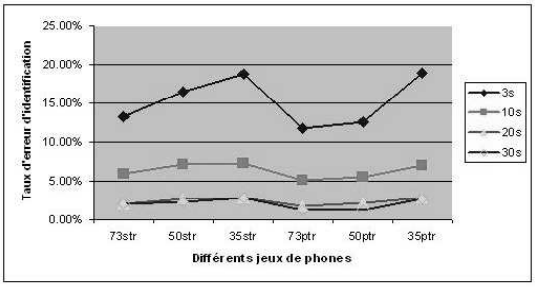


FIG. 3: Taux d'erreur d'identification des systèmes phonotactiques et syllabotactiques (décodage avec treillis). 73ptr signifie le système phonotactique de 73 phones, 73str signifie le système syllabotactique de 73 phones.

3.4. Combinaison de différents décodeurs

La mise en parallèle de plusieurs décodeurs donne une structure PPRLM, qui s'avère très efficace pour diminuer les erreurs. Différents systèmes d'IAL sont mis en œvre : système d'IAL de bi-décodeurs, de tri-décodeurs, quadri-décodeurs et penta-décodeurs. Comme le montre la table 2, la combinaison de plusieurs décodeurs améliore légèrement la performance d'IAL et tend à gommer la différence entre les différents systèmes. En effet pour les monodécodeurs et pour 3 secondes de test, les taux d'erreur d'iden-

tification varient entre 11,7% et 18,9% (13,2% et 18,8%) pour les systèmes phonotactiques avec treillis (syllabotactique avec treillis respectivement). Le système d'IAL qui met quatre décodeurs acoustiques (décodeurs acoustico-phonétiques de 73, 50 et 35 phones, décodeur acoustico-syllabique de 73 phones) en parallèle obtient le meilleur taux d'erreur d'identification de 0,9% sur les segments de 30 secondes. Au-delà, le système d'IAL avec cinq décodeurs n'arrive pas à améliorer davantage les résultats.

4. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté différentes configurations de systèmes d'IAL, faisant toutes appel à des modèles acoustiques multilingues dépendant du contexte. Différents inventaires « phonémiques » multilingues, distinguant un nombre de (classes de) voyelles et de consonnes, varient globalement du simple au double ont été mis à l'épreuve. Des approches phonotactique et syllabotactique ont été utilisées de manière classique (en exploitant la meilleure séquence de phones/syllabes décodée) ou bien en exploitant la méthode à base de treillis, qui permet de mieux exploiter l'information produite lors du décodage phonétique/syllabique. Les inventaires plus grands (50, 73) produisent de meilleurs résultats pour l'approche phonotactique. Globalement l'utilisation de treillis de phones et de syllabes améliore les taux d'identification dans toutes les conditions et pour toutes les durées de test. De manière générale, l'approche syllabotactique n'est pas aussi performante que l'approche phonotactique, mais en situation de combinaison avec l'approche phonotactique elle permet d'obtenir des gains légers dans toutes les configurations impliquant le système syllabotactique à 73 unités de phones. Les expériences avec bi-décodeurs, tri-décodeurs, quadri-décodeurs et penta-décodeurs confirment la fiabilité de la structure PPRLM. Dans ce cadre-là, le système d'IAL fusionnant le décodeur acoustico-phonétique de 73 phones, 50 phones et 35 phones, et le décodeur acoustico-syllabique de 73 phones (figure 1) devient le système le plus performant sur les segments longs (supérieurs à 20s). Pour les travaux futurs, nous prévoyons de tester nos systèmes d'IAL sur plus de langues, et de les examiner sur des corpus de différents types comme la parole spontanée.

TAB. 2: Taux d'erreur d'identification pour 7 langues. 73p : décodage avec treillis, 73s : syllabotactique avec treillis ; la combinaison des systèmes est marqué par +.

Système avec treillis	3s	10s	20s	30s
73p+50p	11,6%	5,4%	2,0%	1,5%
73p+73s	13,9%	6,2%	2,5%	1,5%
35p+35s	15,7%	6,7%	2,7%	1,5%
73s+50s+35s	14,2%	6,4%	2,4%	2,3%
73p+50p+35p	11,6%	5,4%	2,1%	1,5%
73p+50p+35p+73s	11,7%	5,4%	1,8%	0,9%
73p+50p+35p+50s	11,7%	5,7%	2,1%	1,4%
73p+50p+35p+35s	12,0%	5,6%	2,4%	1,5%
73p+50p+35p+73s+35s	12,1%	5,4%	2,1%	1,2%

RÉFÉRENCES

[1] C. Corredor-Ardoy. et al. Language Identification with Language-independent acoustic models. *In Proc. Eurospeech*, Grèce, 1997.

[2] M.A. Zissman. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *In IEEE Trans*, on Speech and Audio Processing 4(1), 1996.

[3] W.M. Cambell. et al. Language Recognition with Support Vector Machine. *In Proc. ODYSSEY*, Espagne, 2002.

[4] M. Adda-Decker. et al. Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification. *In Proc. ICPHS*, Espagne, 2003.

[5] J.L. Gauvain. et al. Language Recognition using Phone Lattices. *In Proc. ICSLP*, Corée, 2004.

[6] N. Thangavelu. et al. Language Identification Using Parallel Syllable-like Unit Recognition. *In Proc. ICASSP*, Canada, 2004.

[7] D. Zhu. et al. Different Size Multilingual Phone Inventories and Context-Dependent Acoustic Models for Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[8] B. Ma. et al. An Acoustic Segment Modeling Approach to Automatic Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[9] P. Matejka. et al. Phonotactic Language Identification using High Quality Phoneme Recognition. *In Proc. Eurospeech*, Portugal, 2005.

[10] J.L. Rouas. Modeling Long and Short-Term Prosody for Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[11] Y. Obuchi. et al. Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization. *In Proc. ICASSP*, Etats-Unis, 2005.

[12] C. White. et al. Discriminative classifiers for language recognition. *In Proc. ICASSP*, Etats-Unis, 2006.

Application des machines à vecteurs support mono-classe à l'indexation en locuteurs de documents audio

Belkacem FERGANI [1,2], *Manuel DAVY* [2] *et Amrane HOUACINE* [1]

[1] LCPTS - USTHB, B.P. 32,El Alia, Bab Ezzouar, Alger, ALGERIE
[2] LAGIS/CNRS, BP 48, Cité Scientifique, 59651 Villeneuve d'Ascq cedex, FRANCE
bfergani2001@yahoo.fr,a_houacine@lycos.fr, Manuel.Davy@ec-lille.fr

ABSTRACT

This paper addresses a new approach based on the Kernel Change Detection algorithm introduced recently by Desobry et al. This new algorithm is applied to the speaker change detection and clustering tasks, which are the key issues in any audio indexing process. We show the efficiency of the method through several experiments using RT'03S NIST data. We discuss also the parameters tuning and compare the results to the well known GLR-BIC algorithm.

1. INTRODUCTION

Avec la multiplication de sources de données multimédia et le développement des techniques de numérisation de l'information, nous assistons à une explosion des bases de données d'archivage. De ce fait se pose un problème crucial : Comment accéder facilement et rapidement à l'information recherchée ? Ces deux critères (rapidité et facilité d'accès) sont incontournables pour toute requête d'utilisateur.

L'indexation en locuteur d'un signal sonore consiste à "structurer" ce signal selon l'information véhiculée par le locuteur. Dans ce contexte, la segmentation en locuteur constitue une étape préalable et déterminante pour la suite du processus d'indexation. Elle consiste à d'abord découper le signal audio en zones homogènes contenant uniquement les informations relatives à un seul locuteur. Cette étape est suivie du regroupement (clustering) de ces segments afin d'assembler les zones appartenant à un seul locuteur.

Ce problème a déjà fait l'objet de nombreuses études (voir par exemple [4, 5] et dans les références qui y sont indiqués. Les techniques standard utilisent généralement des descripteurs acoustiques (souvent des MFCC et leurs dérivées) puis appliquent deux fenêtres d'analyse glissantes sur les données de part et d'autre de l'instant courant. Etant donné les vecteurs acoustiques $\mathbf{x}(n)$, $n = 1, 2, \ldots$, la fenêtre d'analyse située avant l'instant d'analyse n définit l'ensemble passé immédiat $\mathsf{X}_\mathsf{p}(n) = \{\mathbf{x}(n - m_\mathsf{p}), \ldots, \mathbf{x}(n-1)\}$ de m_p vecteurs acoustiques, tandis que l'autre fenêtre contient m_f vecteurs acoustiques représentant l'ensemble futur immédiat $\mathsf{X}_\mathsf{f}(n) = \{\mathbf{x}(n+1), \ldots, \mathbf{x}(n+m_\mathsf{f})\}$. L'objectif de ces techniques classiques est de comparer les ensembles $\mathsf{X}_\mathsf{p}(n)$ et $\mathsf{X}_\mathsf{f}(n)$. Ceci est réalisé au moyen de méthodes à base de rapport de vraisemblance généralisé (RVG) soit directement [5] ou indirectement comme dans l'approche par critère d'information bayésien (BIC) notée RVG-BIC et adoptée comme référence dans ce papier [2]. Les méthodes à base de RVG nécessitent la connaissance d'un modèle de la distribution de probabilité des données $\mathbf{x}(n)$. Les modèles gaussien ou mélange de gaussiennes ont été largement exploités dans ce cadre.

Dans cette communication, nous proposons d'appliquer l'algorithme basé sur une méthode à noyau introduit dans [3] aux tâches de détection de ruptures et de regroupement dont le résultat d'ensemble est connu sous le vocable de segmentation en locuteurs. Différemment des approches citées précédemment, notre algorithme exploite une méthode à base de Machines à Vecteurs de Support mono-classe (SVM-1) dont la finalité est de comparer les ensembles $\mathsf{X}_\mathsf{p}(n)$ et $\mathsf{X}_\mathsf{f}(n)$ à chaque instant d'analyse au moyen d'une mesure de similarité. Dans ce sens, notre méthode reste semblable aux méthodes classiques à base de RVG, néanmoins notre approche exploite les informations extraites de l'entraînement de deux (SVM-1), dont l'avantage principal est de contrôler la complexité du modèle ajusté aux données et de prendre en compte l'information paramétrée selon diverses configurations et tailles des vecteurs acoustiques.

Dans la section suivante, nous rappelons le principe de l'algorithme de détection de rupture basé (SVM-1), puis la section 3 présente la méthodologie d'application à la segmentation en locuteurs en détaillant le choix des paramètres de détection de rupture et de regroupement. La section 4 présente les résultats d'application de notre méthode aux signaux de la base de données NIST RT'03S [6], en comparaison avec la méthode RVG-BIC, selon le critère d'erreur définie par NIST, finalement la dernière section 5 présente les conclusions et perspectives.

2. UN ALGORITHME POUR LA DÉTECTION DE RUPTURES BASÉ SVM-1

Nous partons de l'hypothèse que les vecteurs acoustiques $\mathbf{x}_1, \ldots, \mathbf{x}_m$ sont générées identiquement et indépendemment par une distribution de probabilité (ddp) inconnue $\mathsf{p}(\mathbf{x})$. Le principe de l'algorithme développé dans [3] est de comparer les ensembles $\mathsf{X}_\mathsf{p}(n)$ et $\mathsf{X}_\mathsf{f}(n)$ au travers de la comparaison de leurs support de ddp. On définit le support d'une ddp S^λ par l'ensemble des points de l'espace des vecteurs acoustiques $\mathcal{X}$ telle que $\mathsf{p}(\mathbf{x}) \geq \lambda$, avec λ une constante positive quelconque.

2.1. *Les Machines à Vecteurs Support mono-classe (SVM-1)*

Soit une fonction réelle symétrique appelée noyau définie dans $\mathcal{X}$. Dans la suite, de cette communication, nous considérons un noyau de forme gaussien, comme suit :

$$\mathsf{k}(\mathbf{x}_1, \mathbf{x}_2) \;=\; \exp -\frac{1}{2\sigma^2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2_{\mathcal{X}} \qquad (1)$$

avec $\|\cdot\|^2_{\mathcal{X}}$ est une norme définie sur $\mathcal{X}$. Le modèle SVM-1 estime le support de la ddp comme suit :

$$\mathsf{S}^{\lambda} \;=\; \{\mathbf{x} \in \mathcal{X} | \mathsf{f}^{\lambda}(\mathbf{x}) + b \geq 0\} \qquad (2)$$

Ce problème d'estimation du support de la ddp revient à estimer une fonction dans l'espace augmenté $\mathcal{H}$ (hilbertien et à noyau reproductible induit par $\mathsf{k}(\mathbf{x}_1, \mathbf{x}_2)$) , proche du support de la ddp recherchée. On montre dans [7] que les fonctions minimisant le risque régularisé s'écrivent en fonction de $\mathbf{x}$ comme :

$$\mathsf{f}^{\lambda}(\mathbf{x}) + b \;=\; \sum_{i=1}^{m} \alpha_i \mathsf{k}(\mathbf{x}, \mathbf{x}_i) + b \qquad (3)$$

les coefficients de pondération α_i sont dénommés les multiplicateurs de lagrange. Le paramètre ν (réel positif) joue un rôle de contrôle des vecteurs supports. Ainsi, choisir $\nu = 0.2$ équivaut à admettre 20% de vecteurs acoustiques dans X comme "outliers". A ce problème d'optimisation correpond un problème dual plus simple à résoudre puisque quadratique avec des contraintes linéaires :

$$\begin{array}{ll} \text{Minimiser} & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \text{ w.r.t. } \{\alpha_1, \ldots, \alpha_m\} \\ \text{avec} & 0 \leq \alpha_i \leq \frac{1}{\nu m} \text{ pour } \; i = 1, \ldots, m \\ \text{et} & \sum_{i=1}^{m} \alpha_i = 1 \end{array} \qquad (4)$$

le modèle SVM-1 admet une simple interprétation géométrique dans l'espace augmenté $\mathcal{H}$: premièrement, les vecteurs acoustiques dans $\mathcal{X}$ sont projetés vers $\mathcal{H}$ au moyen de l'application :$\mathbf{x} \rightarrow \mathsf{k}(\mathbf{x}, \cdot)$. Deuxièment, les vecteurs acoustiques dans $\mathcal{H}$ sont de norme unitaire lorsque le noyau gaussien est choisi, car $\|\mathsf{k}(\mathbf{x}, \cdot)\|^2_{\mathcal{H}} = \langle \mathsf{k}(\mathbf{x}, \cdot), \mathsf{k}(\mathbf{x}, \cdot)\rangle_{\mathcal{H}} = \mathsf{k}(\mathbf{x}, \mathbf{x}) = 1$ (propriété du noyau reproductible dans l'espace hilbertien), ansi ces vecteurs sont situés sur la surface d'une hypersphère de rayon unité. Troisièment, la résolution de Eq. (4) peut se ramener à trouver dans $\mathcal{H}$ un hyperplan orthogonal à $\mathsf{f}(\cdot)$ tel que celui-ci serait le plus éloigné de l'origine, séparant ainsi les données d'apprentissage $\mathsf{k}(\mathbf{x}_i, \cdot)$ entre deux classes,— voir figure 1.

2.2. *une mesure de similarité basée sur SVM-1*

Cette mesure est construite sur le principe que les ensembles $\mathsf{X}_\mathsf{p}(n)$ and $\mathsf{X}_\mathsf{f}(n)$ sont similaires si et seulement si les supports de densités estimés sont similaires selon un certain critère de similarité. Notons, que dans l'espace initial des données $\mathcal{X}$, la forme des contours de décision

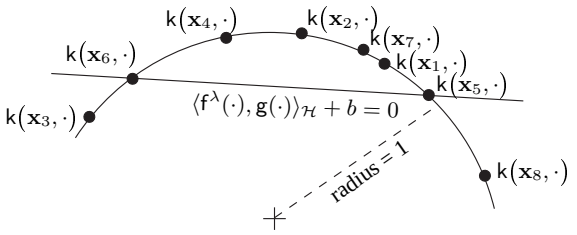


FIG. 1: Interprétation géométrique du modèle SVM-1 dans $\mathcal{H}$. Les vecteurs acoustiques projetés $\mathsf{k}(\mathbf{x}_i, \cdot), i = 1, \ldots, m$ sont situés sur une hypersphère de rayon unité. La fonction $\mathsf{f}^{\lambda}(\cdot)$ et b définissent un hyperplan d'équation $\langle \mathsf{f}^{\lambda}(\cdot), \mathsf{g}(\cdot)\rangle_{\mathcal{H}} + b = 0$. La majorité des données est située du côté de l'hyperplan ne comprenant pas l'origine de l'hypersphère. les coefficients α_i correspondants sont nulles ($i \in \{1, 2, 4, 7\}$), tandis que les points marginaux $\mathsf{k}(\mathbf{x}_3, \cdot)$ et $\mathsf{k}(\mathbf{x}_8, \cdot)$ sont situés du côté de l'hyperplan comprenant l'origine $\alpha_3 = \alpha_8 = 1/\nu m$. Les points situés sur l'intersection de l'yperplan et de l'hypersphère vérifient $0 < \alpha_i < 1/\nu m$ ($i \in \{5, 6\}$).

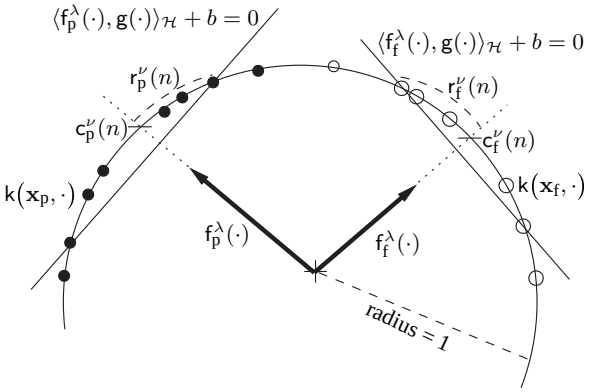


FIG. 2: Interprétation géométrique de l'algorithme basé SVM-1. La situation représentée ici correspond à une détection de ruptures, car les hyperplans représentés par $\mathsf{f}^{\lambda}_\mathsf{p}(\cdot)$ (correspondant à l'ensemble passé immédiat – cercles pleins) et $\mathsf{f}^{\lambda}_\mathsf{f}(\cdot)$ (correspondants à l'ensemble futur immédiat – cercles vides) sont distinctement séparés, et la distance $\mathsf{d}(\mathsf{c}^{\nu}_\mathsf{p}(n), \mathsf{c}^{\nu}_\mathsf{f}(n))$ est grande par rapport aux arcs $\mathsf{r}^{\nu}_\mathsf{p}(n)$ et $\mathsf{r}^{\nu}_\mathsf{f}(n)$.

représentant $\mathsf{S}^{\nu}_\mathsf{p}(n)$ et $\mathsf{S}^{\nu}_\mathsf{f}(n)$ peuvent être complexes et discontinus, rendant ainsi la définition d'une mesure de similarité dans cet espace très difficile. Heureusement, l'interprétation géométrique des SVM-1 dans l'espace augmenté $\mathcal{H}$ permet d'en déduire une mesure très intuitive et simple à mettre en oeuvre : les quantités $\mathsf{S}^{\nu}_\mathsf{p}(n)$ et $\mathsf{S}^{\nu}_\mathsf{f}(n)$ correspondent géométriquement aux hypercercles résultants de l'intersection de l'hypersphère avec l'hyperplan, voir figure 1. Ainsi, la comparaison des quantités $\mathsf{S}^{\nu}_\mathsf{p}(n)$ et $\mathsf{S}^{\nu}_\mathsf{f}(n)$ dans l'espace initial des données $\mathcal{X}$ se ramène à une comparaison dans $\mathcal{H}$ en comparant les hypercercles correspondants dont les centres sont notés $\mathsf{c}^{\nu}_\mathsf{p}(n)$ et $\mathsf{c}^{\nu}_\mathsf{f}(n)$ et les arcs de cercle $\mathsf{r}^{\nu}_\mathsf{p}(n)$ et $\mathsf{r}^{\nu}_\mathsf{f}(n)$, voir figure 2.

La mesure de similarité basée sur notre algorithme est définie comme suit [3] :

$$\mathsf{D}^{\nu}(n) \;=\; \frac{\mathsf{d}\big(\mathsf{c}^{\nu}_\mathsf{p}(n), \mathsf{c}^{\nu}_\mathsf{f}(n)\big)}{\mathsf{r}^{\nu}_\mathsf{p}(n) + \mathsf{r}^{\nu}_\mathsf{f}(n)} \qquad (5)$$

Pratiquement, $\mathsf{D}^{\nu}(n)$ est calculée à partir des coefficients α_i de chaque support de densité $\mathsf{S}^{\nu}_\mathsf{p}(n)$ et $\mathsf{S}^{\nu}_\mathsf{f}(n)$,— voir l'article de F. Desobry et M. Davy dans [3] pour les détails de calcul et de développement de cette mesure.

3. DÉTECTION DE CHANGEMENT DE LOCUTEURS

3.1. Algorithme de détection de ruptures

1. **Parametrisation acoustique** : Celle-ci est effectuée au moyen d'outils conventionnels tel que HTK Tools [8]. Cette étape est effectuée pour l'ensemble du signal.
2. **Entraînement des SVM-1** : Pour chaque instant n et consécutivement à la formation des ensembles $\mathsf{X}_\mathrm{p}(n)$ et $\mathsf{X}_\mathrm{f}(n)$ deux modèles SVM-1 sont entraînés en résolvant le problème (4) pour les deux ensembles. Pour réduire les charges de calculs et accélerer la procédure, la technique développée dans [1] peut être utilisée.
3. **évaluation de la mesure de similarité** : Comme dans toute technique de segmentation (détection de rupture), une rupture est détectée lorsque l'indice $\mathsf{D}^\nu(n)$ définie dans Eq. (5) dépasse un seuil prédéterminé, qui peut être fixe pour tout les instants. Une autre approche consiste à calculer toute la courbe des distances puis choisir un seuil donné permettant d'estimer les instants de ruptures.

3.2. Le regroupement hiérarchique

Suite à la détection de ruptures, nous sommes en présence d'une collection d'objets (segments de paroles homogènes/locuteurs) et nous devons regrouper ces objets par classe (les locuteurs). Nous avons choisi de mettre en oeuvre le regroupement hiérarchique agglomératif qui consiste à considérer au départ chaque segment comme étant une classe et à chaque itération on réunit deux classes les plus proches au sens d'un critère, appelé critère de regroupement [2]. Dans ce cas ce critère est la mesure de similarité définie dans la section 2.2. Ce processus est réitéré jusqu'à l'obtention d'une classe unique. Nous obtenons à l'issue du regroupement un arbre de classification appelé dendrogramme. C'est la manière de parcourir l'arbre qui définit la partition finale.

4. EXPÉRIENCES ET RÉSULTATS

4.1. La Base de Données

Les signaux utilisés sont issus d'enregistrements d'émissions d'informations radio-diffusés (Broadcast News en abrégé bnews) de diverses stations américaines fournies par NIST [6]. Ces fichiers se divisent en deux catégories : 6 fichiers de développement (dry run files) de 10 mn environ chacun dédiés au réglage des paramètres et 3 fichiers d'évaluation (Eval files) de 30 mn chacun.

4.2. Expériences sur les signaux de développement

Afin de régler les paramètres de notre algorithme pour les tâches de détection de ruptures et de regroupement nous utilisons les fichiers de développement pris séparément et le score global est la moyenne globale sur l'ensemble des fichiers. Le critère de performance établi et fourni par l'institut NIST est le "Speaker Diarization Error" ou "Diarrization Error Rate" (DER) fourni par un script en langage perl (voir [4] pour d'amples détails).Pour les expériences concernant cette catégorie de fichiers, la parametrisation utilisée est de 16 coefficients MFCC.

TAB. 1:
Evolution du "DER" en fonction
de la taille des fenêtres glissantes m pour ν=0.1s, σ=0.51 et $\Delta_n = 0.2$s.

m (s)	0.5	0.7	0.9	1.5	2.5	3.5
DER (%)	14.70	13.83	19.21	10.73	16.65	25.36

Sélection de paramètres Les paramètres pertinents de notre algorithme objet de cette étude de sélection sont : $m = m_\mathrm{p}$=m_f (taille de l'ensemble $\mathsf{X}_\mathrm{p}(n)$ et de $\mathsf{X}_\mathrm{f}(n)$ que nous supposons égaux) et le pas de progression Δ_n de ces ensembles appelés communément fenêtres glissantes précédent et succédant à l'instant d'analyse n. Ces deux paramètres m et Δ_n sont communs avec la méthode de référence RVG-BIC. Aux fins de comparaison des deux méthodes nous avons fixé les mêmes valeurs pour ces deux paramètres. Le jeu de paramètres initialement testé dans la table 1 n'est pas fortuit, mais s'inspire des réglages des paramètres de la méthode de référence RVG-BIC, en conformité avec les expériences menées dans [2].

Notre algorithme utilise un paramètre additionnel relatif au noyau, assurant le contrôle de la corrélation des points voisins dans l'espace augmenté $\mathcal{H}$. Dans le cas du noyau gaussien ce paramètre est l'ecart-type σ.

La figure 3 montre l'évolution du DER en fonction de la variation de σ. Le minimum d'erreur est atteint pour σ =0.51, et vaut 10.73%. Les autres paramètres utilisés pour cette expérience sont fixés comme suit : m = m_p=m_f=win=1.5s et Δ_n=0.2s, qui sont des paramètres adéquats pour les deux méthodes en comparaison. Le paramètre ν =0.1 traduit une tolérance maximale de 10% des vecteurs support.

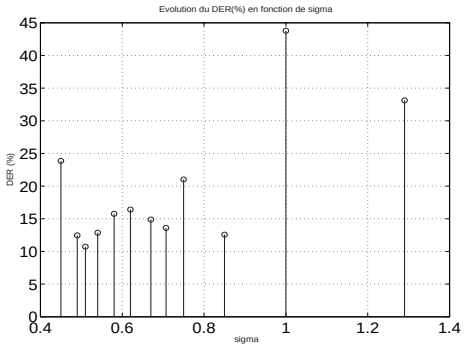


FIG. 3: Evolution du DER en fonction du paramètre du noyau σ.

Les tables 1 et 2 résument les variations de la taille des fenêtres adjacentes glissantes m = m_p=m_f et leurs pas de progression Δ_n. On confirme, que les valeurs de $m = 1.5$s et $\Delta_n = 0.2$s sont des réglages adéquats au regard de la l'erreur obtenue (10.73%).

Evaluations de stratégies de parametrisation acoustique Nous présentons dans cette sous-section l'impact des diverses stratégies de parametrisation acoustiques décrites dans la table 3. Pour chaque configuration, nous

TAB. 2:
Evolution du "DER" en fonction du pas de progression Δ_n, pour $m = 1.5$s, $\nu = 0.1$ et $\sigma = 0.51$.

Δ_n (s)	0.1	0.2	0.3	0.4	0.5	0.6
DER (%)	24.42	10.73	11.93	15.27	12.85	22.68

TAB. 3: Parametrisations acoustiques testées.

Configuration	composition du vecteur acoustique
C_0	16 MFCCs
C_1	16 MFCCs et 10 LPCCs
C_2	C_1 et 10 coefficients de reflection
C_3	C_2 et 10 coefficients banc de filtre
C_4	16 MFCCs et 16 ΔMFCCs
C_5	C_4 et 16 $\Delta\Delta$MFCCs

avons optimisé la sélection des paramètres, tels que mentionnée dans la section ci-dessous. Nous reportons dans les tables 4 et 5 les erreurs minimales obtenues avec le jeu de paramètres sélectionné.

L'estimation du nombre de locuteurs présents dans la conversation analysée est obtenue en parcourant le dendrogramme selon une coupe horizontale, ce qui revient à faire une hypothèse sur le nombre de locuteurs désiré puis de vérifier celle-ci selon la performance obtenue. Les travaux de synthèse reportés dans [4] offrent une explication claire sur la détermination automatique du nombre de locuteurs.

4.3. Validation sur les fichiers d'évaluation

La table 6 montre que notre méthode obtient des taux d'erreurs DER bien inférieurs à la méthode de référence RVG-BIC. Le meilleur résultat obtenu est la moyenne sur ces trois fichiers, soit un taux d'erreur de 13.63%. Ces résultats sont d'autant plus prometteurs car comparés à ceux publiés récemment dans la littérature constituant l'état de l'art [4] (page 22,Table 7) et dans laquelle les méthodes présentées ont été optimisées indépendamment de notre algorithme.

5. CONCLUSIONS ET PERSPECTIVES

Les résultats présentés dans cette communication montrent clairement que notre approche basée sur un algorithme à base des machines à vecteurs support monoclasse ouvre une voie de recherche très prometteuse pour l'indexation en locuteurs de discours multi-locuteurs. Nous imputons cette performance, à un meilleur processus de segmentation acoustique plutôt qu'à un meilleur regroupement, du fait que c'est la première phase qui conditionne la seconde, néanmoins, une affirmation rigoureuse ne peut découler que d'une étude comparative détaillée de la pureté moyenne des segments obtenues à la suite du processus de regroupement. Ce travail est une perspective déja entammée. Un autre résultat intéressant concerne l'étude comparée des méthodes RVG-BIC et SVM-1 en fonction des diverses stratégies de parametrisation. Ainsi, il apparaît, qu'une paramétrisation même redondante améliore les résultats globalement pour les deux méthodes mais que c'est notre méthode qui assure une nette supériorité.

TAB. 4: Performances comparées (KCD/RVG) en fonction des parametrisations acoustiques décrites dans Table 3.

Config	DERmin (%)		estim. # loc.	
	KCD	RVG-BIC	KCD	RVG-BIC
C_0	10.73	26.38	17	9
C_1	8.37	21.18	17	9
C_2	7.95	15.08	17	11
C_3	10.90	15.26	9	9
C_4	11.44	20.91	13	11
C_5	8.63	14.30	19	11

TAB. 5: Jeu de paramètres en fonction des parametrisations acoustiques testes.

Configuration	Jeu de paramètres selectionné
C_0	$m = 1.5$s,$\sigma = 0.51$, $\Delta_n = 0.2$s.
C_1	$m = 2.0$s,$\sigma = 1$, $\Delta_n = 0.3$s
C_2	$m = 1.5$s,$\sigma = 1$, $\Delta_n = 0.3$s
C_3	$m = 2.5$s,$\sigma = 0.707$, $\Delta_n = 0.2$s
C_4	$m = 0.7$s,$\sigma = 0.707$, $\Delta_n = 0.2$s
C_5	$m = 0.9$s,$\sigma = 0.85$, $\Delta_n = 0.3$s

RÉFÉRENCES

[1] M. Davy, F. Desobry, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. *Signal Processing*, 2006. to appear.

[2] P. Delacourt and C. Wellekens. Distbic : a speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1) :111–126, September 2000.

[3] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Trans. Sig. Proc.*, 53(5), August 2005.

[4] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language* , pages 1–28, 2005. in Press.

[5] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In *IEEE ICASSP'04*, Montreal, Canada, 2004.

[6] NIST RT03S. The rich transcription spring 2003 (rt-03s) evaluation plan http ://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt-03-spring-eval-plan-v4.pdf,. 2003.

[7] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, USA, 2002.

[8] S. Young and all. *The HTK Book (for HTK Version 3.2.1)*. cambridge.

TAB. 6: Résultats obtenus sur les signaux d'évaluation. Les paramètres sont $m = 1.5$s, $\nu = 0.1$, $\sigma = 0.51$ et $\Delta_n = 0.2$s. La parametrisation choisie est C_2. Les fichiers sont (a) 20010228.2100-2200-MNB-NBW, (b) 20010217.1000-1030-VOA-ENG and (c) 20010220.2000-2100-PRI-TWD.

Fichier	RVG-BIC	KCD	Estimation du Nbre de Locuteurs
(a)	25.60	14.34	23
(b)	20.17	12.28	25
(c)	22.69	14.27	22

Indexation en locuteur : utilisation d'informations lexicales

J. Mauclair, S. Meignier, Y. Estève

LIUM, Université du Maine Le Mans, France
{julie.mauclair,sylvain.meignier,yannick.esteve}@lium.univ-lemans.fr

ABSTRACT

The automatic speaker indexing consists in splitting the signal into homogeneous segments and clustering them by speakers. However the speaker segments are specified with anonymous labels. This paper propose to identify those speakers by extracting their full names pronounced in the show. With a semantic classification tree, the full names detected in the segment transcription are associated to this segment or to one of its neighbors. Then, a merging method associates a full name to a speaker cluster instead of the anonymous label. The experiments are carried out over French broadcast news from the ESTER 2005 evaluation campaign. About 70% show duration is correctly processed for evaluation corpus.

1 Introduction

Les transcriptions manuelles d'enregistrements audio sont très coûteuses, particulièrement lorsqu'on cherche à indexer des informations spécifiques telles que le thème principal, les mots-clés, le nom du locuteur... Seules les méthodes automatiques génèrent des transcriptions enrichies à moindre coût, mais leur taux d'erreur doit être suffisamment faible pour pouvoir les exploiter. Ici, nous nous intéressons uniquement au problème de l'identité du locuteur.
La première étape pour obtenir automatiquement des transcriptions enrichies consiste à segmenter le signal puis à regrouper les segments en locuteur. Les principales méthodes reposent uniquement sur des paramètres acoustiques [1, 2]. L'étape suivante consiste à transcrire automatiquement les segments résultants.
L'indexation attribue uniquement des étiquettes anonymes aux segments, alors que connaître la véritable identité du locuteur serait judicieux pour la recherche documentaire de documents sonores. Les méthodes existantes permettant d'associer le nom complet (prénom, nom) d'un locuteur aux segments issus de l'indexation sont de deux sortes. Les premières méthodes reposent sur le traitement d'informations purement acoustiques, avec généralement l'utilisation d'un système de reconnaissance automatique du locuteur requérant des échantillons de voix pour apprendre les modèles [3]. Les autres reposent sur des informations lexicales et extraient l'identité du locuteur directement à partir la transcription de l'émission. Le nom du locuteur et sa localisation sont présents dans les mots prononcés durant une émission de radio et ces informations peuvent être utilisées pour identifier le locuteur avec son véritable nom. Aucun échantillon de voix n'est ici nécessaire. En effet, les intervenants se présentent ou présentent l'intervenant suivant, ils félicitent le précédant ou le suivant, concluent le reportage par leur nom ... Dans de récents travaux menés sur des émissions radiophoniques en anglais [4], le LIMSI utilise des règles linguistiques extraites manuellement pour identifier le locuteur du segment avec son véritable nom. En fonction des annonces de qui parle, de qui parlera ou de qui vient de parler, un nom détecté dans la transcription permet d'étiqueter avec la véritable identité du locuteur le segment courant, le suivant ou le précédent. Le taux d'erreur de ce processus basé sur des règles manuelles est d'environ 13% à partir de transcriptions manuelles (18% à partir de transcriptions automatiques).
Dans ce papier, nous proposons une association automatique du locuteur avec son nom complet grâce à l'utilisation d'un arbre de classification sémantique qui apprend automatiquement ce genre de règles. Cette méthode fournit seulement une décision locale pour le segment courant et les segments contigüs. L'identité du locuteur est ensuite propagée sur la totalité de l'émission de radio.
Pour réaliser l'étude préliminaire présentée ici, nous utilisons en entrée pour le système les indexations en locuteur ainsi que les transcriptions manuelles de référence.

Nous gardons à l'esprit que les erreurs provenant de l'indexation et de la transcription automatique réduisent les performances (voir résultats de [4]). Les corpus proviennent de la campagne d'évaluation française ESTER [5]. Cependant, le processus entièrement automatique utilisé nous permet d'adapter facilement la méthode à d'autre langues.

2 Informations sur le locuteur

2.1 Identité cliente

Les participant à des émissions radiophoniques sont principalement des personnes publiques comme des journalistes, des politiciens, des artistes ou des sportifs. Cette population est facilement identifiable : leurs noms et prénoms sont bien connus, ils sont présents dans plusieurs émissions, et ils correspondent aux locuteurs principaux en termes de temps de parole. Ces locuteurs sont identifiés par la concaténation de leurs noms et prénoms dans les transcriptions d'ESTER. Ce sont les locuteurs à identifier dans la tâche proposée (locuteurs clients). 1007 noms complets différents ont été extraits des corpus utilisés dans nos expériences en ne conservant que les noms des personnes publiques. Le procédé de détection du nom du locuteur repose sur cette liste fermée. Nous avons choisi d'employer le nom complet pour éviter les fausses détections introduites par la méthode de détection : l'ambiguïté présentée par l'utilisation de noms partiels (seulement le prénom ou le nom) amène des problèmes que nous ne résoudrons pas ici.

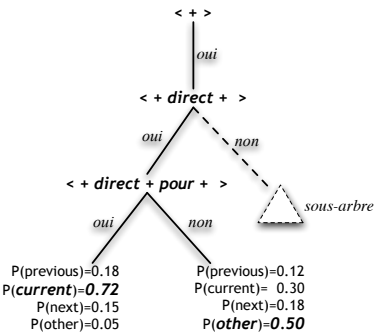


Figure 2: Exemple d'une partie d'un arbre de classification sémantique : à chaque feuille, une probabilité est associée à chaque étiquette.

2.2 Étiquetage des occurrences de noms

Un nom complet détecté dans un segment peut être associé à une des 4 étiquettes suivantes : *current*, *next*, *previous* et *other*. Elles sont attribuées respectivement si le nom détecté semble se rapporter au locuteur du segment de parole courant, du segment suivant ou du segment précédent (voir figure 1). Si ce n'est pas le cas, l'étiquette *other* est attribuée au nom détecté.

3 Méthode employée

Pour associer un nom complet à une étiquette anonyme issue de l'indexation en locuteur (figure 1, partie ①), nous proposons les deux étapes suivantes :
1- **Analyse du contexte lexical pour chaque un nom complet** (figure 1, partie ②) : cette étape traite chaque nom complet détecté dans la transcription d'un segment de parole. Elle détermine si ce nom se rapporte au locuteur précédent, courant, suivant ou à un autre locuteur. Seuls les segments proches d'un nom détecté dans la transcription peuvent être associés à ce nom. D'ailleurs, des segments peuvent être associés à différents noms : les processus d'association sont faits sans coopération et peuvent fournir des résultats antagonistes sur un même segment.
2- **Dénomination du locuteur** (figure 1, partie ③) : la deuxième étape consiste à fusionner les hypothèses précédentes afin d'associer un nom complet à une étiquette anonyme d'un locuteur et de répercuter ce nom à tous les segments étiquetés avec cette même étiquette anonyme. (figure 1, partie ④).

3.1 Analyse du contexte lexical

Quand un nom est détecté, le contexte lexical de la transcription est analysé pour associer à ce nom l'étiquette la plus pertinente. Cette analyse est faite en employant un arbre de classification sémantique (SCT) [6].

Arbre de classification sémantique Les SCTs reposent sur l'utilisation d'expressions régulières. Des couples composés d'une occurrence de nom complet et de son contexte lexical sont classés en utilisant ces expressions régulières. Notre but est de classer ces couples parmi les 4 étiquettes *previous*, *current*, *next* et *other* Pour un nom complet et son contexte lexical, chaque feuille de l'arbre peut donner une probabilité pour chaque étiquette possible. La figure 2 donne un exemple d'expression régulière utilisée pour construire l'arbre.

Décisions locales Pour une occurrence o de nom complet détectée dans un contexte lexical $W_s(o)$ (voir section 4.1) associé à un segment de parole s, un SCT peut donner la probabilité $P(t|W_s(o))$ pour chaque étiquette possible t de l'ensemble des étiquettes $T = \{previous, current, next, other\}$. Soit l'étiquette $\delta(o) \in T$ associée à une occurrence d'un nom complet du segment de parole s telle que :

$$\delta(o) = \underset{t}{\operatorname{argmax}}\, P(t|W_s(o))$$

Dans notre approche, parmi les 4 étiquettes possibles pour $W_s(o)$, seule l'étiquette $\delta(o)$ est prise en considération pour la suite du processus. En outre, si plus d'une étiquette obtient une probabilité égale à $\max_t P(t|W_s(o))$, aucune décision locale n'est prise.

Définissons la valeur $\Gamma(o)$ qui servira par la suite telle que $\Gamma(o) = P(\delta(o)|W_s(o))$.

3.2 Dénomination du locuteur

Soit ψ un locuteur anonyme d'un segment de parole : il s'agit de trouver le vrai nom $N(\psi)$ de ce locuteur. Chaque segment de parole est associé à un locuteur anonyme (par exemple dans la figure 1, le segment 1 est associé à SPK1, ainsi que les segments 9 et 11). De plus, en utilisant un arbre de classification sémantique sur les noms complets détectés dans la transcription des segments, on obtient une liste de noms correspondant aux locuteurs possibles pour quelques segments (figure 1, partie ②).

Fusion des décisions prises par le SCT Soit K, l'ensemble de tous les noms complets des locuteurs clients. Soit ν_ψ, l'ensemble des différents noms complets associés à au moins un segment prononcé par ψ grâce à une décision locale du SCT : ν_ψ est la liste des noms complets candidats pour ψ et on a $\nu_\psi \subset K$. Soit la fonction $\nu(o)$ qui associe une occurrence de nom complet o à ce nom complet n. Enfin, soit l'ensemble Ω_ψ des occurrences o qui réfèrent aux segments prononcés par ψ grâce aux décisions locales prises par le SCT.
Pour trouver le nom complet $N(\psi)$ d'un locuteur ψ, nous proposons la formule suivante :

$$N(\psi) = \underset{n\in K}{\operatorname{argmax}} \frac{\sum_{(\nu(o)=n)\wedge(o\in\Omega_\psi)} \Gamma(o)}{\sum_{o\in\Omega_\psi} \Gamma(o)} \quad (1)$$

$$= \underset{n\in K}{\operatorname{argmax}} \sum_{(\nu(o)=n)\wedge(o\in\Omega_\psi)} \Gamma(o) \quad (2)$$

Ainsi, le nom complet qui sera associé à une étiquette anonyme de locuteur est le nom dont les occurrences maximisent la somme des scores données par le SCT quand ces occurrences se rapportent à des segments associés à cette étiquette. Comme expliqué dans la section 3.1, seules les valeurs associées aux décisions locales valides sont conservées.

4 Expériences et résultats

4.1 Données

Corpora Les méthodes proposées sont développées et évaluées avec les données de la campagne ESTER 2005. ESTER est une campagne d'évaluation sur les systèmes de transcription d'émissions radiophoniques en Français [5]. Les données (Table 1) comportent six radios différentes dont les émissions durent de 10 à 60 min et sont décomposées en 3 corpora. Le corpus d'apprentissage (*Train*) contient 81h de données

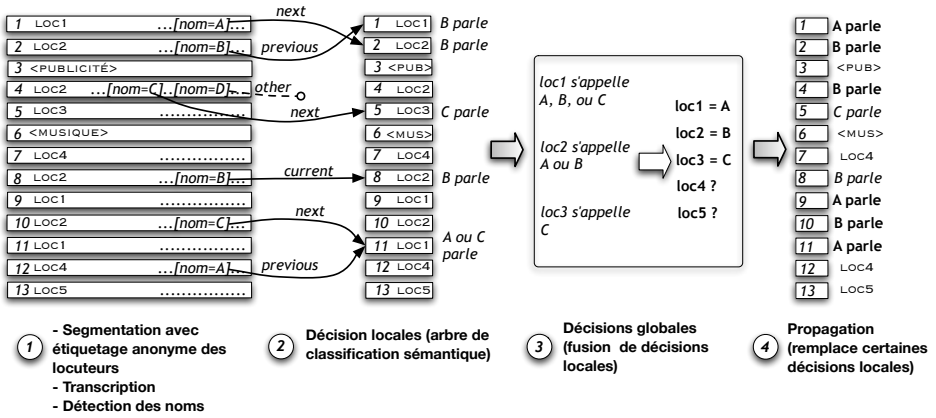


Figure 1: Identification du locuteur

Table 1: Détails sur les corpus : Apprentissage, Développement & Évaluation provenant de la campagne d'évaluation ESTER et statistiques sur les ≠ étiquettes applicables aux noms.

	Train	***Dev***	***Eva***
# Radios	5	5	6
durée (h)	81	12.5	10
# Segments	8547	2294	1417
Previous (%)	14.3	12.6	18.6
Current (%)	7.2	7.1	5.3
Next (%)	46.0	45.3	49.3
Other (%)	32.5	35.0	26.8

(8547 segments) dans lesquels 3297 noms complets sont détectés. Le corpus de développement (*Dev*) [1] (*Dev*) contient 12,5h (2294 segments) et 920 noms complets. Le corpus d'évaluation (*Eva*) contient 10h (1417 segments) et 507 noms complets, il correspond au corpus d'évaluation officiel d'ESTER.Le tableau 1 montre également les probabilités *a priori* des 4 étiquettes sur ces corpus.

Préparation des différents corpus Les références (transcriptions enrichies) doivent être transformées et adaptées pour être utilisées par un arbre de classification sémantique et également pour évaluer les résultats expérimentaux. Les adaptations effectuées sont :
• La définition des 4 étiquettes de noms complets suppose que les locuteurs voisins sont différents du locuteur courant. Les segments contigus contenant le même locuteur sont donc fusionnés pour obtenir une segmentation basée sur les changements de locuteur.
• Les informations concernant les 4 étiquettes doivent être accessibles durant les différentes phases. Nous avons donc étiqueté automatiquement la référence en extrayant du flux audio les noms et prénoms des locuteurs. Chaque nom complet extrait est comparé au nom de locuteur associé au segment ainsi qu'aux noms de locuteurs associés aux segments voisins. Cette tâche étant automatisée, nous supposons qu'il n'y a pas d'erreur d'identification du locuteur.
• Pour généraliser les exemples d'apprentissage pour la construction de l'arbre, chaque nom de locuteur est remplacé par une étiquette générique.
• Le SCT apprend les expressions régulières en tenant compte des mots appartenant aux contextes gauche et droit de l'occurrence du nom détecté. Au plus 20 mots pour le contexte gauche et 20 pour le droit sont conservés, ce nombre ayant été fixé sur le corpus de développement *Dev* pour maximiser le nombre de bonnes détections locales sur les 4 étiquettes.

[1] fusion des corpus de développement officiels des phase I et II d'ESTER

4.2 Etiquetage des segments

Table 2: Scores des décisions locales obtenus grâce au SCT sur les différents corpus.
- Etiquetés : % de noms complets détectés pour lesquels une règle de décision locale prorpose un étiquetage other, current, previous ou next.
- Previous (resp. pour chacune des étiquettes) : % de noms complets détectés correctement étiquetés par previous.

	Train	***Dev***	***Eva***
# Noms complets détectés	3297	920	507
Etiquetés (%)	94.51	94.78	97.23
Correctement Etiquetés (%)	88.25	76.49	68.76
Previous (%)	88.98	71.67	82.98
Current (%)	94.76	90.14	85.71
Next (%)	89.32	80.67	75.29
Other (%)	84.87	68.94	50.32

L'arbre de classification sémantique qui donne les résultats du tableau 2 a été construit à partir du corpus d'apprentissage. Le tableau montre les résultats des décisions locales prises sur chaque segment contenant un nom complet sur les corpus *Train*, *Dev* et *Eva*. La première colonne montre les résultats sur les données d'apprentissage utilisées comme des données de test.

94% des noms complets détectés sur *Dev* et 97% sur *Eva* sont associés à une de nos 4 étiquettes. Le taux d'étiquetage correct est d'environ 76.4% sur *Dev* et de seulement 68.7% sur *Eva* : ces valeurs peuvent être considérées comme étant la précision de la décision locale sur chaque corpus.
Le taux moins élevé de*Eva* (environ 8% de moins que *Dev*) peut être expliqué par la présence de deux nouvelles radios non présentes dans les autres corpus et ces données ont en plus été enregistrées 15 mois après les autres. Environ 6% des noms détectés ne sont pas étiquetés comme pour le corpus d'apprentissage.

Les résultats pour l'étiquette *other* sont les plus mauvais. Cela peut s'expliquer par le fait que cette étiquette est confrontée à une plus grande diversité de contextes lexicaux que les autres. Néanmoins, puisque l'étiquette *other* n'intervient pas directement

dans le processus global de prise de décision, l'impact de ces mauvais résultats est minime.

En choisissant toujours l'étiquette ayant la plus forte probabilité *a priori* (tab. 1), le meilleur score serait d'environ 49.3% d'étiquetage correct sur le corpus *Eva*. Avec la méthode proposée, nous atteignons le score d' environ 68% d'étiquetage correct pour *Eva*. Ces résultats montrent que l'utilisation d'un arbre de classification sémantique s'insère efficacement dans un processus de dénomination du locuteur.

4.3 Dénomination du locuteur

Table 3: Dénomination du locuteur : résultats détaillés pour les différents corpus (les taux sont calculés en termes de durée). *- Loc. : correpond aux 2 catégories de locuteurs de la référence, ceux qui sont les locuteurs clients (C) de l'application (locuteurs publiques avec un nom complet) et les autres (NC).*
- Dénom. : correspond aux dénominations correctes et incorrectes. "Non nommé" correspond au cas où le processus ne propose pas de nom.

Loc.	Dénom.	***Train***	***Dev***	***Eva***
C	Cor.(%)	63.68	64.82	66.35
C	Inc.(%)	3.19	5.48	14.36
C	Non nommé(%)	15.68	18.19	11.91
NC	Cor.(non nommé)(%)	15.50	7.54	3.59
NC	Inc.(%)	1.95	3.98	3.79
	Total(%)	100	100	100

Les décisions locales sur les segments sont ensuite fusionnées pour associer un nom complet à tous les segments de parole prononcés par le même locuteur (voir section 3.2). Les résultats détaillés de cette seconde étape sont reportées dans le tableau 3.

Méthode d'évaluation L'entrée du système est basée sur les transcriptions manuelles de référence : il n'y a ni d'erreurs d'indexation , ni de segmentation parole/non parole ni de transcriptions. Les frontières des segments de référence et d'hypothèse sont les mêmes, seuls les noms de leurs locuteurs diffèrent. Ici, seuls les locuteurs qui sont des personnes publiques (avec un nom complet dans la référence) sont les locuteurs clients. L'identité des autres ne peut pas être trouvée. Il y a donc erreur quand le processus donne à un locuteur non-client un nom complet et quand il ne donne pas de nom à un locuteur client (Table 3 lignes 2 & 5). De plus, le processus ne propose pas de nom à un locuteur client dans les cas où :
• aucune décision locale n'atteint un segment de ce locuteur client. Soit aucune décision n'est prise pour toutes les occurrences détectées de ce locuteur, soit les décisions locales attribuées sont fausses ;
• ce nom n'est pas détecté dans la transcription.
Pour les locuteurs clients, quand les noms complets hypothèses et ceux de référence sont les mêmes, la dénomination est correcte (Table 3 ligne 1). Pour les locuteurs non-clients, quand le processus ne propose pas de noms, il semble raisonnable de considérer cela comme correct (Table 3 ligne 4).
Tous les résultats proposés sont calculés en terme de durée comme c'est le cas lors des évaluations NIST [7].

Commentaires Le processus de dénomination de locuteur atteint 72% de décision correcte en terme de durée (64.82% + 7.54%) sur le corpus *Dev* et environ 70% (66.35% + 3.59%) sur *Eva* (voir tableau 3).

5 Conclusion

Dans le contexte de la transcription enrichie, nous proposons une méthode totalement automatisée qui permet d'identifier les locuteurs par leur véritable identité extraite directement de la transcription.
Le procédé proposé est basé sur l'utilisation d'un arbre de classification sémantique qui permet d'étiqueter les occurrences de noms détectées : cette première étape consiste à prendre des décisions locales qui associent une telle occurrence à un segment de parole. Ensuite, les résultats obtenus sont fusionnés pour associer un nom complet à tous les segments d'un même locuteur qui était anonymement annoté par l'indexation en locuteur.
Les expériences sont menées sur des émissions radiophoniques Françaises, fournies par la campagne d'évaluation ESTER. Environ 70% de la durée totale des émissions est correctement indexée en locuteur pour chacun des corpus de développement et d'évaluation. Sur le corpus d'évaluation, 18,15% de la durée totale des émissions est incorrectement indexée et aucune décision n'est prise pour 11.91%.
Le but principal est atteint : les résultats valident la méthode proposée de dénomination du locuteur à partir d'une indexation et d'une transcription manuelle. Les perspectives de travail s'orientent vers l'utilisation d'une indexation en locuteur et d'une transcription automatiques dans lesquelles des erreurs interviendront.

Remerciements

Merci à Frédéric Béchet du LIA pour la mise à disposition sous licence GPL de LIA_SCT. LIA_SCT est un classifieur basé sur un arbre de décision disponible sur le site web du LIA.

Bibliographie

[1] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *DARPA RT04 Fall*, Palisades, NY, USA, 2004.

[2] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, 2005.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, vol. 4, pp. 430–451, 2004.

[4] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," Jeju, Oct 2005.

[5] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of European Conference on Speech Communication and Technology – Interspeech 2005*, Lisboa, Sep 2005, pp. 1149–1152.

[6] R. Kuhn and R. De Mori, "The application of semantic classifciation trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 449–460, 1995.

[7] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," `http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v%14.pdf`, August 2004.

Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique

E. Didiot, I. Illina, O. Mella, D. Fohr, J.-P. Haton

LORIA-CNRS & INRIA Lorraine
BP 239, 54506 Vandoeuvre-les-Nancy, France
Mél : {didiot,illina,mella,fohr,jph}@loria.fr

ABSTRACT

The problem of Speech/Music discrimination is a challenging research problem which significantly impacts Automatic Speech Recognition (ASR) performance. This paper proposes new features for the Speech/Music discrimination task. We use a decomposition of the audio signal based on wavelets which allows a good analysis of non stationary signals like speech or music. We compute different energy types in each frequency band obtained from wavelet decomposition. We use two Class/Non-Class classifiers : one for speech/non speech, one for music/non music. On a broadcast corpus, using the proposed wavelet approach, we obtained a significant improvement (35%) compared to MFCC parameters.

1. INTRODUCTION

La discrimination entre la parole et la musique consiste à segmenter le flux audio en zones acoustiquement homogènes comme la parole, la musique ou la parole sur fond musical. C'est un problème important pour l'indexation de documents audio et pour la transcription automatique de programmes radiophoniques. Dans le cadre de la transcription, la séparation entre parole et musique permet d'éliminer les segments ne contenant que de la musique et donc de diminuer le nombre d'erreurs de reconnaissance.
La discrimination parole/musique nécessite deux étapes : la paramétrisation puis la classification du signal audio. L'étape de paramétrisation consiste à extraire du signal des caractéristiques discriminantes entre la parole et la musique. De nombreux paramétrages ont été proposés dans l'état de l'art du domaine. Ils peuvent être classés en paramétrages fréquentiels, temporels, mixtes et dans le domaine cepstral.
Les paramètres fréquentiels sont calculés à partir de la DSP (Densité Spectrale de Puissance). La DSP d'un signal est issue de la transformée de Fourier de la fonction d'auto-corrélation. Les paramètres fréquentiels permettent de capter le contenu fréquentiel d'un signal à un moment donné (formants, harmoniques, etc.). Le centroïde spectral, le flux spectral et le *Spectral Rolloff Point* [12, 13] sont les plus utilisés.
Les paramètres temporels sont calculés directement à partir du signal audio. Ils permettent de capter des éléments relatifs à la variation de l'onde acoustique, comme par exemple, la périodicité ou les variations de l'onde sonore au cours du temps. Parmi ces paramètres, citons la mesure de rythmicité (*Pulse Metric*), l'énergie et le taux de passage par zéro (*Zero Crossing Rate*) [12, 13, 14].
Les paramètres mixtes proviennent à la fois d'analyses fréquentielle et temporelle du signal. La modulation d'énergie à 4Hz [13], la modulation basse fréquence de l'amplitude du signal dans différentes bandes de fréquence de Karneback [6] ou encore le pourcentage de trames de faible énergie [12] en sont des exemples.
Les Coefficients Cepstraux (MFCC) permettent une représentation compacte du spectre du signal en prenant en compte, grâce à l'échelle Mel, des informations perceptives. Ces paramètres, très utilisés en reconnaissance de la parole, sont utilisés non seulement en discrimination parole/musique [1, 7, 10] mais aussi en classification de genres musicaux [15].
Cet article présente une nouvelle approche de discrimination parole/musique fondée sur l'utilisation de la décomposition en ondelettes du signal. A notre connaissance, une telle approche n'a jamais été utilisée pour cette tâche. Notre motivation pour l'utilisation des ondelettes est qu'elles mesurent les variations temporelles des composantes spectrales et permettent donc d'extraire les caractéristiques temporelles et fréquentielles du signal. De plus, la décomposition multi-bandes, qu'effectue la transformée en ondelettes dyadique, se rapproche fortement de ce que fait l'oreille humaine [4], c'est à dire qu'elle ressemble à une échelle logarithmique. Par rapport aux coefficients MFCC, les ondelettes ont une résolution temps-fréquence différente, permettent une représentation plus compacte du signal, possèdent un ensemble plus riche de fonctions de base et sont plus robustes à la non stationnarité du signal et à ses distorsions.
Dans notre travail, nous avons étudié différentes décompositions en ondelettes du signal, en extrayant des paramètres fondés sur l'énergie. Pour valider l'approche proposée, nous avons effectué des expériences sur un corpus d'émissions radiophoniques. Nous avons comparé la paramètrisation proposée au paramètrage MFCC.
Les sections 2 et 3 introduisent nos nouveaux paramètres et décrivent notre système de classification parole/musique. La section 4 présente les corpus utilisés tandis que les résultats de nos expériences sont détaillés dans la section 5. Enfin, la section 6 présente nos conclusions.

2. PARAMÈTRES FONDÉS SUR LES ONDELETTES

L'approche fondée sur les ondelettes est une approche de traitement du signal. Elle a été utilisée avec succès pour une large variété de problèmes, comme par exemple le débruitage ou récemment en reconnaissance automatique de la parole [11].
La transformée en ondelettes discrète (*Discrete Wavelet Transform*, DWT), que nous utilisons, analyse le si-

gnal dans différentes bandes de fréquence à différentes résolutions. Une telle analyse permet de contrôler les variables temps et fréquence du signal. Stéphane Mallat [8] a montré qu'une telle décomposition peut être obtenue par des filtrages passe-bas et passe-haut du signal temporel. Après chaque filtrage le signal est sous-échantillonné d'un facteur deux. Ce processus de décomposition est itéré sur les résultats du filtrage passe-bas jusqu'à l'obtention du nombre de bandes de fréquence désiré. Au final, le signal est décomposé en coefficients d'*approximation* et en coefficients de *détail*. Les coefficients d'approximation correspondent à des moyennes locales du signal. Les coefficients de détail, appelés « coefficients d'ondelettes », représentent les différences entre deux moyennes locales successives, c'est à dire entre deux approximations successives du signal.
Dans notre étude, nous utilisons la transformée en ondelettes dyadique correspondant à un banc de filtre en bandes d'octave. Dans le cas discret, cette transformée en ondelettes peut être définie par l'équation suivante :

$$W(k,j) = \sum_j \sum_k x(n) 2^{\frac{-j}{2}} \Psi(2^{-j}n - k)$$

où $x(n)$ correspond au signal à l'instant n, $\Psi()$ est une fonction temporelle de moyenne nulle, d'énergie finie et à décroissance rapide, appelée « ondelette mère ». La transformée en ondelettes *dyadique* effectue une décomposition du signal en bandes de fréquence non uniformes, ce qui permet d'obtenir une résolution fréquentielle décroissante lorsque les fréquences augmentent. Ainsi, cette décomposition en ondelettes donne une analyse multi-résolution du signal : une haute résolution temporelle et une basse résolution fréquentielle dans les hautes fréquences et inversement dans les basses fréquences. Elle offre ainsi une bonne modélisation du système auditif humain. Dans cet article, nous utilisons uniquement les coefficients d'ondelettes pour paramétrer le signal acoustique. L'utilisation des coefficients d'ondelettes permet de capter les modifications brutales du signal. En effet, les coefficients d'ondelettes prennent une grande valeur lors de tels événements.
Dans chaque bande de fréquence, nous calculons à partir des coefficients d'ondelettes, différents paramètres d'énergie [2].
Nous calculons :

– *Logarithme de l'énergie* (**E**) : l'énergie instantanée :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j}(w_k^j)^2\right)$$

où w_k^j dénote le coefficient d'ondelettes à la position k et à l'échelle j, N_j le nombre de coefficients à l'échelle j, et f_j le vecteur de paramètres à l'échelle j.

– *Logarithme de l'énergie Teager* (**T_E**) : nous utilisons ici l'opérateur TEO (*Teager Energy Operator*) introduit par Kaiser [5]. Cet opérateur permet d'obtenir des paramètres robustes :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j-1}|(w_k^j)^2 - w_{k-1}^j w_{k+1}^j|\right) \quad (1)$$

– *Logarithme de l'énergie hiérarchique* (**H_E**) : l'énergie hiérarchique correspond au calcul de l'énergie au centre de la fenêtre d'analyse en utilisant le même nombre de coefficients quelque soit la bande :

$$f_j = \log_{10}\left(\frac{1}{N_J}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2}(w_k^j)^2\right)$$

où J correspond à la résolution la plus basse. Cette énergie a été utilisée avec succès en paramétrisation pour la reconnaissance automatique de la parole [3].

3. SYSTÈME DE DISCRIMINATION PAROLE/MUSIQUE

3.1. Paramètrisation

Le signal est échantillonné à 16kHz. Après préaccentuation, nous utilisons une fenêtre de Hamming de 32ms avec un déplacement de 10ms. Nous avons comme paramètres :
Paramètres MFCC de référence : 12 coefficients MFCC avec leur première et seconde dérivées (36 coefficients).
Paramètres fondés sur les ondelettes : ces paramètres sont calculés en utilisant deux familles d'ondelettes : daubechies et coiflet. Les paramètres multi-résolutions sont calculés pour différents niveaux de décompositions, c'est-à-dire différents nombres de bandes (de 5 à 7).

3.2. Description du système

Pour classifier les segments audio nous avons utilisé l'approche « anti-modèles »[9] : les modèles de classe et de non classe sont utilisés en parallèle pendant la classification. Chaque classe est modèlisée par un GMM dont le nombre de gaussiennes est compris entre 8 et 64. Deux sous-systèmes sont mis en oeuvre : parole/non parole et musique/non musique.
Après le regroupement des décisions de ces deux sous-systèmes, le signal audio est classé en 3 catégories : parole (P), musique (M) et parole sur fond musical (PM). Pour éviter les segments de très courte durée, nous avons imposé une durée minimale de 0.5s pour chaque segment reconnu : pour cela nous concatenons 50 GMMs pour former un modèle HMM. Pour trouver la meilleure séquence de modèles qui décrit le signal, l'algorithme de Viterbi est utilisé.

4. CORPUS

4.1. Corpus d'apprentissage

Nous avons entraîné nos modèles sur deux corpus : *CDs audio* et *programmes radio*. Le corpus *CDs audio* (120 mn) est constitué de morceaux de musique instrumentale et de chansons, extraits de CDs. Le corpus *programmes radio* (976 mn) contient des programmes de radios françaises : des journaux, des interviews et des programmes musicaux. Ces programmes sont très variés en terme de parole et de musique, de styles d'élocution, de locuteurs, de conditions d'enregistrement (téléphone, studio, interviews, bruits).

4.2. Corpus de test

Nous avons effectué nos expériences sur un corpus composé de 3 parties :

– La partie *News* est composée de trois fichiers d'une heure de bulletins d'information de radios françaises (« France-Inter »et « Radio France International ») et

contient principalement de la parole.
- La partie *Entertainment* est composée de trois émissions de 20 minutes chacune (interviews et programmes musicaux). Cette partie est considérée comme difficile. En effet, elle comporte beaucoup de segments superposés (parole avec musique ou chanson) et des effets de *fade in-fade out*. Cette partie comprend aussi une alternance de parole de qualité studio et de qualité téléphonique. De plus, certaines interviews sont très bruitées.
- La partie *Scheirer* correspond au corpus de test construit et utilisé par E. Scheirer et M. Slaney [13]. Tous les fichiers audio sont homogènes et ont la même durée de 15 secondes : 20 fichiers de parole studio ou téléphonique et 41 fichiers de musique ou de voix chantée. Les styles musicaux sont plus nombreux (jazz, pop, country, etc.) que dans la partie *Entertainment*. Cette partie ne contient pas de parole sur fond musical. Lors du test, le système n'a pas connaissance que les fichiers de ce corpus sont homogènes, et il peut donc commettre des erreurs de segmentation.

Au total, ce corpus de test contient 74% de parole, 12% de parole sur musique et 14% de musique.

5. RÉSULTATS EXPÉRIMENTAUX

Pour évaluer notre système, nous avons utilisé 3 scores. Soit n_z^y le nombre de trames reconnues comme z alors qu'elles étaient étiquetées y, et soit nT, le nombre total de trames. Nous calculons :
- Taux de classification correct Global (TG) :

$$100 * (n_{PM}^{PM} + n_M^M + n_P^P)/nT$$

- Taux de classification Musique/Non Musique (M/NM) :

$$100 * (n_{PM}^{M} + n_M^{PM} + n_M^M + n_{PM}^{PM} + n_P^P)/nT$$

- Taux de classification Parole/Non Parole (P/NP) :

$$100 * (n_{PM}^{P} + n_P^{PM} + n_M^M + n_{PM}^{PM} + n_P^P)/nT$$

Pour le système de référence avec les coefficients MFCC, nous obtenons les résultats suivants : **TG = 82,0%, M/NM = 84,1% et P/NP = 96,2%**. L'intervalle de confiance est de $0,5\%$ à 5% de risque.

5.1. *Influence du niveaux de décomposition et du calcul de l'énergie*

Nous avons évalué des paramètrisations fondées sur différentes énergies (instantanée (**E**), Teager (**T_E**), hiérarchique (**H_E**)) calculées à partir des coefficients d'ondelettes. Après des expériences préliminaires, nous avons choisi d'utiliser les ondelettes suivantes : daubechies à 4 moments nuls (*db-4*) et coiflet à 2 moments nuls (*coif-1*). Différents niveaux de décomposition (nombre de bandes de fréquence) sont testés : de 5 à 7.

La table 1 montre que les meilleurs résultats en M/NM et P/NP sont obtenus avec l'ondelette *coif-1* : pour M/NM avec l'énergie hiérarchique sur 7 bandes et pour P/NP avec l'énergie Teager sur 5 bandes. Ainsi, en M/NM nous obtenons un gain relatif significatif de 42% (gain absolu de 6,7%) par rapport aux MFCC. En P/NP les résultats obtenus sont comparables aux MFCC. Notons que l'énergie hiérarchique permet une meilleure discrimination en M/NM lorsque le nombre de niveau de décomposition augmente. Le meilleur taux global est obtenu avec l'ondelette *coif-1* et l'énergie Teager sur 7 bandes : nous obtenons une diminution significative du taux d'erreurs de 6,7% comparé aux MFCC, soit un gain absolu de 1,2%.

Dans la table 1, la deuxième colonne correspond au nombre de bandes de fréquence utilisé pour la décomposition en ondelettes et donc au nombre de paramètres. Nous observons que nos paramètres offrent une représentation plus compacte du signal que les MFCC : avec seulement 7 coefficients, les paramètres basés sur les ondelettes sont plus performants que les MFCC avec 36 coefficients. Enfin, pour la discrimination parole/non-parole, nous obtenons avec l'énergie *Teager* des résultats comparable aux MFCC, mais avec moins de coefficients. Cette bonne performance de l'énergie Teager s'explique peut-être par le fait que l'opérateur de Teager permet de prendre en compte la dynamique à très court terme du signal (cf équation 1).

5.2. *Influence des paramètres dynamiques*

La durée d'une trame (32ms) n'est pas suffisante à un être humain pour faire la différence entre de la parole et de la musique. E. Scheirer et M. Slaney ont montré que l'utilisation de la variance sur une seconde de leurs paramètres améliorait les résultats en discrimination parole/musique[13]. Ainsi, l'étude de paramètres à plus long terme semble être intéressante.

Pour étudier la dynamique de nos paramètres à moyen terme, nous avons ajouté à nos paramètres statiques leurs dérivées première (Δ) et seconde ($\Delta\Delta$) (voir table 2). Pour étudier la dynamique à plus long terme, la variance des paramètres statiques, calculée sur une fenêtre d'une seconde, est utilisée à la place des paramètres eux-mêmes (voir table 3). Nous avons choisi comme paramètres statiques les paramètres qui ont donnés les meilleurs résultats auparavant : *coif-1* avec 7 bandes de fréquence.

La table 2 montre que l'ajout des dérivées a permis d'améliorer la discrimination parole/musique (TG) de 24% (avec Δ) et de 16% (avec $\Delta\Delta$) par rapport aux paramètres MFCC. Toutefois, on constate que l'ajout des dérivées secondes n'a pas apporté d'amélioration par rap-

TAB. 1: Influence du niveau de décomposition et du type d'énergie en utilisant les ondelettes *db-4* et *coif-1*.

Type ond.	Bds	Ener.	M/NM	P/NP	TG
MFCC+Δ+ΔΔ			*84,1*	***96,2***	*82,0*
db-4	5	E	84,2	95,4	81,2
db-4	5	T_E	84,9	95,2	81,4
db-4	5	H_E	84,8	83,3	78,2
db-4	6	E	86,2	93,5	82,2
db-4	6	T_E	86,1	94,6	82,9
db-4	6	H_E	87,7	91,2	81,7
db-4	7	E	82,2	93,7	78,7
db-4	7	T_E	83,0	94,7	80,1
db-4	7	H_E	88,0	89,1	81,6
coif-1	5	E	88,5	95,2	82,0
coif-1	5	T_E	88,6	**96,0**	82,0
coif-1	5	H_E	84,7	82,9	76,8
coif-1	6	E	86,1	93,8	81,1
coif-1	6	T_E	87,5	95,2	82,6
coif-1	6	H_E	90,7	89,0	82,4
coif-1	7	E	88,0	92,4	82,1
coif-1	7	T_E	88,7	93,2	**83,2**
coif-1	7	H_E	**90,8**	85,7	80,1

TAB. 2: Influence des paramètres dynamiques (Δ et $\Delta\Delta$) en utilisant l'ondelette *coif-1* avec 7 bandes.

Param.	Nb	M/NM	P/NP	TG
MFCC+Δ+ΔΔ	36	*84,1*	*96,2*	*82,0*
E+Δ	14	88,0	95,9	85,9
T_E+Δ	14	88,3	**96,2**	**86,3**
H_E+Δ	14	**88,6**	95,8	84,1
E+Δ+ΔΔ	21	88,0	95,8	84,7
T_E+Δ+ΔΔ	21	86,8	96,1	84,9
H_E+Δ+ΔΔ	21	84,3	95,6	82,2

TAB. 3: Influence des paramètres dynamiques (variance sur 1 seconde) en utilisant l'ondelette *coif-1* avec 7 bandes.

Param.	Nb	M/NM	P/NP	TG
MFCC+Δ+ΔΔ	36	*84,1*	*96,2*	*82,0*
MFCC+Δ+ΔΔ(Var1s)	36	*86,4*	*95,9*	*84,2*
E Var 1s	7	90,4	96,0	88,0
T_E Var 1s	7	**90,6**	**96,4**	**88,4**
H_E Var 1s	7	90,3	88,7	84,1

port à l'ajout des dérivés premières. En ce qui concerne l'utilisation de la variance calculée sur une durée d'une seconde, elle permet d'obtenir de meilleurs résultats à la fois pour nos paramètres et pour les paramètres MFCC (table 3). Par exemple, la variance calculée sur les paramètres *énergie Teager* a permis d'améliorer significativement le Taux Global de 35% par rapport aux MFCC. Enfin, notons que nous obtenons ces derniers bons résultats en utilisant seulement 7 coefficients par vecteur de paramètres.

5.3. Expérimentations sur la partie Scheirer

Pour comparer nos résultats avec ceux obtenus par Scheirer et Slaney [13], nous avons utilisé le même corpus de test. Notons que le corpus d'apprentissage diffère de celui de Scheirer et Slaney. Nous comparons notre parametrisation avec la paramètrisation de Scheirer et Slaney suivante : la modulation de l'énergie à 4Hz, la variance du flux spectral et la mesure de rythmicité (*Best 3*). La classification est faite trame par trame, sans contrainte de durée minimum sur les segments. La table 4 montre que nos résultats sont comparables à ceux publiés dans [13].

TAB. 4: Performance par trame (%) sur la partie *Scheirer*.

Param.	M/NM	P/NP
MFCC+Δ+ΔΔ	*88,9*	*94,6*
MFCC+Δ+ΔΔ (Var 1s)	*91,4*	*95,4*
Scheirer (Best 3)	*93,6*	***95,8***
coif-1, 7 bds, T_E (Var 1s)	**94,1**	94,9

6. CONCLUSION

Dans cet article, nous avons proposé de nouveaux paramètres fondés sur la décomposition en ondelettes du signal audio et sur le calcul de différentes énergies. Ces paramètres sont utilisés pour la tâche de discrimination parole/musique. Par rapport aux MFCC, la décomposition en ondelettes fournit une résolution temporelle non uniforme pour les différentes bandes de fréquence. Cette paramétrisation permet également d'obtenir une représentation plus compacte du signal et d'être plus robuste à la non-stationnarité des signaux. La décomposition dyadique que nous effectuons nous fournit une approximation de l'échelle Mel. Les paramètres finaux sont obtenus en calculant différents types d'énergie sur les coefficients d'ondelettes.
Notre nouvelle paramétrisation donne de meilleurs résultats de discrimination parole/musique que les coefficients MFCC avec leurs paramètres dynamiques. Nous avons ainsi un gain relatif significatif de 42% en classification musique/non musique et 6,7% en classification globale. L'utilisation des paramètres dynamiques (dérivées premières et seconde, variance sur 1s) améliore encore plus nos résultats : un gain relatif significatif de 35% en discrimination parole/musique, 3% en P/NP et 40% en M/NM, est obtenu par rapport à la paramétrisation MFCC. Notons que notre paramétrisation utilise un nombre réduit de coefficients : 7 coefficients par rapport aux 36 coefficients MFCC. Plusieurs perspectives sont envisageables. D'une part, une fusion de différentes paramétrisations nous semble intéressante (MFCC+ondelettes). D'autre part, étant donné que la variance sur 1s a montré de bons résultats, l'étude d'autres paramètres à long terme peut s'avérer prometteuse.

7. REMERCIEMENTS

Nous remercions Eric Scheirer et Malcolm Slaney pour nous avoir fourni leur corpus de parole et de musique.

RÉFÉRENCES

[1] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A Comparison Of Features For Speech, Music Discrimination. In *ICASSP-99*, 1999.

[2] M. Deviren. *Systèmes de reconnaissance de la parole revisités : Réseaux Bayesiens dynamiques et nouveaux paradigmes*. PhD thesis, Université Henri Poincaré, 2004.

[3] R. Gemello, D. Albesano, L. Moisa, and R. De Mori. Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System. In *ICASSP-01*, 2001.

[4] S. Maes I. Daubechies. A Nonlinear Sqeezing of The Continuous Wavelet Transform based on Auditory Nerve Models. In *Wavelets in Medecine and Biology*, 1996.

[5] J.F. Kaiser. On a Simple Algorithm to Calculate the 'Energy'of a Signal. In *ICASSP-90*, 1990.

[6] S. Karneback. Discrimination between Speech and Music based on a Low Frequency Modulation Feature. In *European Conf. on Speech Comm. and Technology*, 2001.

[7] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[8] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[9] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle*. PhD thesis, Université Paul Sabatier (Toulouse III), 2004.

[10] J. Razik, D. Fohr, O. Mella, and N. Parlangeau-Vallès. Segmentation Parole/Musique pour la transcription automatique. In *JEP04*, 2004.

[11] R. Sarikaya and J.H.L. Hansen. High Resolution Speech Feature Parameterization for Monophone-based Stressed Speech Recognition. *IEEE Signal Processing Letters*, 7(7) :182–185, 2000.

[12] J. Saunders. Real-Time Discrimination of Broadcast Speech/Music. In *ICASSP-96*, 1996.

[13] E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *ICASSP-97*, 1997.

[14] C.C.J. Kuo T. Zhang. Hierarchical System for Content-Based Audio Classification and Retrieval. In *Conference on Multimedia Storage and Archiving Systems III, tome 3527 de SPIE*, 1998.

[15] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Transaction on Speech and Audio Processing*, 10(5) :293–302, 2002.

Représentation paramétrique des relations temporelles appliquée à l'analyse de données audio pour la mise en évidence de zones de parole conversationnelle

Zein Al Abidin IBRAHIM, Isabelle FERRANÉ, Philippe JOLY

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 31062 Toulouse cedex
{ibrahim, ferrane, joly}@irit.fr

ABSTRACT

The aim of our work is the automatic analysis of audiovisual documents to characterize their structure. Our approach is based on the study of the temporal relations between events occurring in a same document. For this purpose, we have proposed a parametric representation of temporal relations, from which Temporal Relation Matrix can be computed and analyzed to identify relevant relation classes. In this paper, we applied our method on audio data, mainly on speaker and applause segmentations from a TV game program. Our purpose is to analyze these basic audio events, to see if the observations automatically highlighted could reveal information of a higher level like speaker exchanges or conversation, which may be relevant in a structuring or indexing process.

1. INTRODUCTION

Analyser automatiquement des documents audiovisuels en vue d'en dégager la structure ou de les caractériser de façon à pouvoir les rattacher à des catégories de documents est l'une des principales motivations de nos travaux. L'indexation automatique de documents audiovisuels passe d'abord par le développement d'outils dont le rôle est de détecter des caractéristiques primaires sonores ou visuelles (parole, musique, bruit, couleur, quantité de mouvement). Les résultats produits sont généralement des séquences de segments dans lesquels la caractéristique primaire recherchée est ou non présente. Ces informations bas niveau doivent ensuite servir de base à la recherche d'événements plus pertinents, utilisés pour la génération de résumé ou l'indexation en vue d'effectuer des traitements plus évolués.

Dans les documents audiovisuels, le temps est une composante essentielle. Notre approche est basée sur l'analyse des relations temporelles qui peuvent être observées entre événements présents dans un document. Compte tenu de l'aspect multimédia des données que nous manipulons, nous nous sommes fixé pour objectif de proposer une méthode générique, qui puisse utiliser tout type d'informations, sonores ou visuelles, primaires ou de niveau un peu plus élevé (locuteur, visage, plan, …). Cette méthode d'analyse sera décrite dans la section 2.

Des travaux récents sur les documents audiovisuels ont porté sur la détection ([6] et références associées) et l'analyse de scènes conversationnelles [2]. D'abord basés sur l'étude de la composante vidéo, couleur, plans (segmentation, durée, similarité), détection et identification de visages, certaines caractéristiques audio (énergie, zones de parole) peuvent être utilisées pour renforcer la détection de conversations. Des caractéristiques audio seules, détection de zones voisées et zones de parole [2] peuvent également être utilisées. Comme dans l'ensemble de ces travaux, la parole conversationnelle, dans sa forme plutôt que dans la teneur des échanges entre individus, nous intéresse particulièrement. En effet, la notion de conversation en tant que successions de tour de parole est un type d'événements qui est indissociable de l'aspect temporel. Notre motivation, dans la présentation de cet article, est donc d'étudier comment et en quoi notre méthode peut aider à la détection de zones conversationnelles dans un document audiovisuel. Bien que notre méthode soit définie pour être indépendante du type de caractéristique utilisé, pour cette étude, nous nous focaliserons sur des données uniquement audio et présenterons les résultats de notre analyse dans la section 3. La section 4 présentera la conclusion de ce travail et les perspectives de travaux futurs.

2. REPRÉSENTATION PARAMÉTRIQUE DES RELATIONS TEMPORELLES

L'analyse temporelle du contenu de documents audiovisuels, élargie à tout type de relations temporelles observables entre tout type d'événements, nécessite de disposer d'outils pour représenter et raisonner sur des informations temporelles. Deux types de modèles temporels existent : ceux utilisant le point comme unité temporelle [5] et ceux basés sur la notion d'intervalle temporel [1]. Cette seconde représentation est la plus adaptée à nos travaux qui reposent sur l'utilisation de segmentations élémentaires d'un même document.

2.1. Segmentation élémentaire

On définit une segmentation élémentaire comme un ensemble de N intervalles temporels disjoints correspondant à un seul et même type d'événements

survenant dans le document traité : présence d'une caractéristique audio (applaudissement, parole, locuteur) ou visuelle (personne visible à l'écran). Dans la mesure où des outils de segmentation automatique sont disponibles, nous en utilisons les résultats, sinon, les segmentations sont produites manuellement. Soit **S** une telle segmentation qu'on notera $\mathbf{S} = \{ \mathbf{s_i} \}$ avec $i \in [1,N]$, $\mathbf{s_i}$ un segment représenté par ses deux extrémités : début $(\mathbf{s_{id}})$ et fin $(\mathbf{s_{if}})$ et noté $\mathbf{s_i} = [\mathbf{s_{id}}, \mathbf{s_{if}}]$.

2.2. Relation temporelle entre deux segments

Soient deux segmentations élémentaires S1 et S2 réalisées sur un même document et telles que :

$$\mathbf{S1} = \{ s_{1i} \}\ i \in [1,N_1] \text{ et } \mathbf{S2} = \{ s_{2j} \}\ j \in [1,N_2]$$
$$\text{avec } \mathbf{s_{1i}} = [s_{1id}, s_{1if}] \text{ et } \mathbf{s_{2j}} = [s_{2jd}, s_{2jf}].$$

La relation temporelle observable entre un couple de segment $(\mathbf{s_{1i}}, \mathbf{s_{2j}})$ est alors représentée par trois variables [4] : $\mathbf{DE} = s_{2jf} - s_{1if}$ $\mathbf{DB} = s_{1id} - s_{2jd}$ $\mathbf{Lap} = s_{2jd} - s_{1if}$

Ceci peut également être formulé de la façon suivante : $\mathbf{s_{1i}}$ **R (DE, DB, Lap)** $\mathbf{s_{2j}}$ où **R** est la relation observée entre les deux segments $\mathbf{s_{1i}}$ et $\mathbf{s_{2j}}$ avec pour paramètres DE, DB et Lap. En considérant chaque triplet comme les coordonnées d'un point dans un espace à trois dimensions, on obtient une représentation graphique des relations temporelles, comme illustré plus loin.

2.3. Matrice des relations temporelles (TRM)

Si on généralise cela à tous les couples de segments issus des deux segmentations considérées, et que l'on associe chaque paramètre à une dimension, on peut alors créer une matrice tridimensionnelle dans laquelle chaque élément est considéré comme un compteur et comptabiliser ainsi toutes les occurrences d'une même relation. Toutes les relations observables sont alors déduites de la comparaison deux à deux des segments de S1 avec les segments de S2. Cet histogramme tridimensionnel permet d'étudier les fréquences des relations temporelles observables. La matrice peut également être utilisée comme matrice de votes, leur distribution permettant d'identifier des règles générales relatives au comportement temporel des événements présents dans le document. Avant d'étudier le contenu de la TRM, quelques petites transformations sont nécessaires. Elles concernent l'uniformisation des unités de temps qui peuvent ne pas être les mêmes d'une segmentation à l'autre lorsque celles-ci ne concernent pas le même média. Elles concernent également les valeurs utilisées comme index pour accéder à un élément de la matrice. Au départ réelle, leur valeur doit être ramenée à une valeur entière ce qui permet de réduire la taille de la matrice.

2.4. Classification des données de la TRM

Contrairement aux techniques de vote classiques, identifier une relation temporelle pertinente ne peut pas se limiter à la recherche d'un maximum local, mais plutôt à celle d'une zone dans laquelle les votes seraient majoritairement distribués. Cela revient à localiser des nuages de points dans la représentation graphique ou de votes dans la TRM. Cela nous conduit soit à procéder à une étape de classification du contenu de la TRM ou bien à décomposer la matrice en régions différentes en fonction de connaissances a priori sur la nature des relations que l'on souhaite observer, comme les relations de Allen [1] par exemple. En effet, la définition même des relations de Allen introduit des contraintes qui limitent la zone représentative de la relation. La table 1 représente les contraintes imposées par les relations temporelles de Allen *'avant'* et *'après'*.

Table 1: Représentation paramétrique des relations 'avant' et 'après'

Relation	DE	DB	Lap
avant	DE > Lap	DB < -Lap	0 < Lap = a
après	DE < 0	DB > 0	DE-DB < Lap < 0 et 0<DB –DE+Lap = a

Lap mesure l'écart entre les deux segments. Si $\mathbf{s_{1i}}$ est avant $\mathbf{s_{2j}}$ alors **Lap** doit être strictement positif. De plus si on veut que les relations observées restent significatives, alors comparer deux segments lorsqu'ils sont trop éloignés l'un de l'autre peut ne pas avoir de sens. Une seconde limite nommée α est alors introduite. Dans la figure 1 ci-dessous on constate que la zone relative à la relation *'après'* est limitée dans l'espace tridimensionnel considéré. Les contraintes relatives aux autres relations de Allen sont données dans [3].

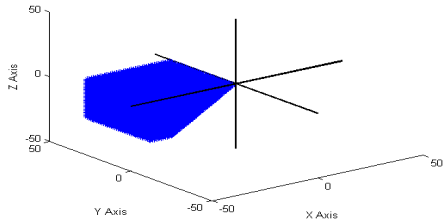


Figure 1 : Représentation graphique de la relation *'après'* avec x = DE, y = DB , z = Lap

Une fois les zones correspondant aux classes de relations identifiées, on leur associe un nombre d'occurrences correspondant à la somme des votes qu'elles contiennent. Ainsi la taille de la matrice est réduite à une matrice cubique, chaque dimension correspondant au nombre de classes ou de relations prédéfinies considérées.

2.5. Combinaison de relations temporelles

Chaque classe de relations temporelles peut correspondre à un type d'événements. Combiner les relations temporelles appartenant à des classes

différentes peut être un moyen de mettre en évidence des événements sémantiquement plus significatifs quant au contenu du document.

La conjonction de relations temporelles a déjà été étudiée dans [1]. Notre approche permet de procéder à la conjonction de relations temporelles et de produire un résultat sous la même forme paramétrique. Soient trois intervalles temporels, $\mathbf{s_{1i}}$, $\mathbf{s_{2j}}$ et $\mathbf{s_{3k}}$ appartenant respectivement à trois segmentations **S1, S2 S3**, effectuées sur le même document, deux relations temporelles **R1** et **R2** peuvent être définies par :

$\mathbf{s_{1i}}$ **R1** (a_1, b_1, c_1) $\mathbf{s_{2j}}$ et $\mathbf{s_{2j}}$ **R2** (a_2, b_2, c_2) $\mathbf{s_{3k}}$

Une nouvelle relation temporelle **R3** résultant de la conjonction des relations **R1** et **R2** pourra être exprimée par : $\mathbf{s_{1i}}$ **R3** (a_3, b_3, c_3) $\mathbf{s_{3k}}$

avec $a_3 = \mathbf{a_1 + a_2}$; $\mathbf{b_3 = b_1 + b_2}$; $\mathbf{c_3 = c_1 - b_2}$

Si **R1** appartient à la classe de relation **C1**, **R2** à la classe de relation **C2** alors une nouvelle classe de relations temporelles **C3** pourra être mise en évidence par l'ensemble des conjonctions pouvant être calculées entre les relations de la première classe et celles de la seconde. Ainsi en procédant hiérarchiquement on peut espérer faire remonter des informations pertinentes et de plus haut niveau sémantique que celles utilisées en entrée du traitement et mieux caractériser le contenu du document traité. Nous avons mené une étude en ce sens pour voir si l'information que l'on peut récupérer est significative et exploitable. Cette étude est présentée dans la section suivante.

3. ETUDE POUR LA MISE EN EVIDENCE DE ZONES DE PAROLE CONVERSASIONNELLE

3.1. Contexte expérimental de l'étude

Notre méthode se veut avant tout la plus générique possible, vis-à-vis du type du document traité, comme des relations temporelles observées ou des événements mis en évidence. Pour cet article nous avons utilisé des données audio et notamment des segmentations élémentaires en locuteur effectuées manuellement sur une émission de jeu télévisée d'une durée de trente et une minutes. Huit locuteurs y sont présents et par conséquent, huit segmentations élémentaires, numérotées de 1 à 8 ici, servent de base à cette étude.

3.2. Calcul des TRM par couple de locuteurs

Suivant le principe de construction présenté dans la section 2.3, une TRM a été calculée pour chaque couple de locuteurs distincts, soit 28 TRM. Ces TRM vont être traitées pour voir si l'aspect conversationnel peut être mis en évidence par l'analyse des relations temporelles entre segments. Comme spécifié dans la section 2.4, il est nécessaire de définir des contraintes pour limiter le champ des observations à réaliser. La contrainte α notamment, introduit une limite dans l'écart des couples de segments à comparer. Initialement fixée à 1, soit une seconde d'écart maximum entre deux segments, nous avons obtenu des résultats très pauvres. La limite α a été augmentée ($\alpha = 10$) pour tenir compte en réalité, des dix secondes de réflexion qui peuvent séparer deux échanges. Cette contrainte est donc, malheureusement, influencée par la nature du document. Pour illustrer le type de résultats que nous avons obtenu, considérons les segmentations numéro 2, 3, 4 et 5 ainsi que les TRM associées. La représentation graphique des $TRM_{2,3}$ et $TRM_{4,5}$ est donnée ci-dessous, figures 2.a et 2.b, les classes C1 et C2 représentées illustrant la partie 3.3.

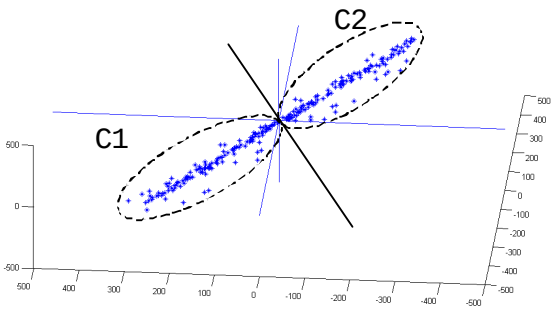


Figure 2.a : Représentation graphique de la $TRM_{2,3}$

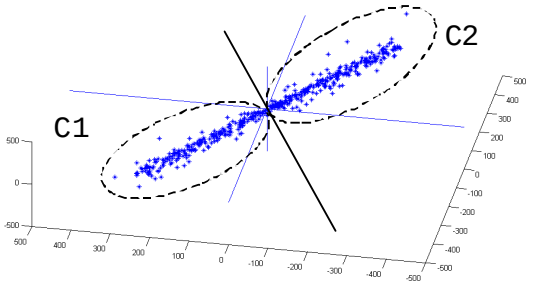


Figure 2.b : Représentation graphique de la $TRM_{4,5}$

Les $TRM_{2,4}$, $TRM_{2,5}$, $TRM_{3,4}$, $TRM_{3,5}$ sont quasiment vides car comparativement, le nombre global de votes pour les relations temporelles observées dans chaque cas est très faible, alors qu'il atteint 247 pour la $TRM_{2,3}$ et 450 pour la $TRM_{4,5}$. En regardant rétrospectivement le contenu du document étudié, et plus particulièrement les locuteurs concernés, on constate que les couples (2,3) et (4,5) correspondent respectivement aux deux équipes impliquées dans le jeu. La nature du jeu veut qu'il y ait plus d'interactions intra équipes qu'inter équipe. Ceci se reflète dans les TRM. Trois autres observations peuvent être faites d'après l'étude du contenu des TRM. Le locuteur 1 est le seul locuteur à interagir avec chacun des autres locuteurs. Il s'avère que ce locuteur est l'animateur du jeu et son rôle veut qu'il s'adresse à tous. La deuxième observation montre qu'au contraire le locuteur 6 n'est en relation qu'avec le locuteur 1 (cf. observation précédente) et le locuteur 4. On constate effectivement que dans le jeu, le locuteur 6 correspond en réalité à une personne du public avec lequel un des joueurs (le 4), joue une manche spéciale. Enfin, les données des TRM indiquent un nombre plus élevé de votes entre les locuteurs de la seconde équipe

(4,5) que ceux de le première (2,3) (cf. plus haut). En effet, une manche consiste pour le joueur d'une équipe à faire deviner à son équipier autant de mots ou d'expression qu'il le peut dans le temps imparti. La seconde équipe est effectivement la meilleure et la plus rapide, ce qui peut justifier le nombre plus élevé d'échanges détectés.

3.3. Analyse des TRM et classification en deux classes.

Pour aller plus loin dans l'analyse des TRM, nous avons opéré une classification de chacune d'elles en deux classes en utilisant la méthode des k-means. La table 2 présente pour chaque $TRM_{A,B}$ les nombres de votes répartis sur les deux classes C1 et C2.

Table 2 : Classification des TRM en 2 classes

A B	**C1**	**C2**	**A B**	**C1**	**C2**	**A B**	**C1**	**C2**
1 2	65	60	**1 3**	49	49	**1 4**	84	71
1 5	106	97	**1 6**	6	5	**1 7**	89	79
1 8	3	5	**2 3**	*123*	*124*	**2 4**	4	7
2 5	6	6	**2 6**	0	0	**2 7**	6	7
2 8	0	0	**3 4**	6	5	**3 5**	10	5
3 6	0	0	**3 7**	7	4	**3 8**	0	0
4 5	*245*	*205*	**4 6**	4	8	**4 7**	15	19
4 8	0	0	**5 6**	0	0	**5 7**	39	26
5 8	0	0	**6 7**	1	0	**6 8**	0	3
7 8	4	3						

Une interprétation possible de la répartition en deux classes, peut être faite en considérant que lorsqu'on examine les locuteurs deux à deux, on a les cas de figure suivants : soit ils ne se parlent pas et ne sont pas en relation (pas de vote enregistré), soit A parle à B (C1), ou B parle à A (C2). La répartition des votes donnée ici, ne donne qu'une vision globale des échanges entre locuteurs au cours de l'émission. Le recours à l'opération de conjonction doit peut être permettre de détecter les zones d'échanges consécutifs.

3.4. Conjonction des relations temporelles

Une conversation peut être considérée ici comme une suite d'échanges entre deux locuteurs, où *A parle à B* (A/B) puis *B parle à A* (B/A), ainsi de suite. Faire la conjonction de relations temporelles de la classe C1 avec celles de la classe C2 (resp. C2 et C1) peut-être un moyen de faire apparaître le motif : (A/B/A) (resp. (B/A/B). Par ce biais, nous avons pu identifier la plupart des suites d'échanges entre deux locuteurs, les plus longues étant pour les couples (2,3) et (4,5). Nous avons ensuite introduit une nouvelle segmentation manuelle, indiquant la présence d'applaudissements. Des TRM associant chaque locuteur (loc_X) à ce nouvel événement (app) ont été calculées et classifiées en 2 classes correspondant aux motifs (loc_X/app) et (app/loc_X). En généralisant l'application de l'opération de conjonction entre les classes appropriées, on peut faire apparaître le motif (A/B/app) ou (A/B/A/B/app) etc. En effet, un coup réussi dans le jeu correspond à une suite d'échanges entre joueurs de la même équipe suivie d'applaudissements. Ces motifs, répétés à nouveau pourront s'ils sont détectés, correspondre à une manche du jeu. Par contre lorsqu'un coup échoue, aucun applaudissement ne suit, et notre méthode fusionne directement avec le coup suivant. La durée des applaudissements pourrait permettre de distinguer entre le gain d'un coup lors d'une manche et le gain de la manche elle même.

4. CONCLUSION ET PERSPECTIVES

Nous avons étudié, à travers un exemple, comment la représentation paramétrique des relations temporelles entre événements peut conduire, après plusieurs phases d'analyse, à mettre en évidence des événements d'une teneur plus sémantique que les événements utilisés à la base. A partir d'informations sur les locuteurs présents dans un même document, l'analyse et la classification des TRM ont fait remonter des informations qui peuvent caractériser les zones conversationnelles et être des indicateurs non seulement de la structure du document mais également du rôle de certains locuteurs et de leurs interactions. Par cette étude nous avons voulu montrer quelle était la nature des résultats que nous pouvions obtenir et quelles étaient les pistes intéressantes à exploiter pour nos travaux futurs. La détermination du nombre optimal de classes dans l'étape de classification et l'application de notre méthode à d'autres catégories de documents comportant d'autres types d'échanges (débats, interviews, variétés) restent encore à évaluer.

BIBLIOGRAPHIE

[1] J. F. Allen. Maintaining Knowledge about Temporal Intervals. *In Communication of ACM*, vol. 26, n°11, pp. 832 – 843, 1983.

[2] S. Basu. Conversational Scene Analysis. *Ph.D. Thesis. MIT Department of EECS*. September, 2002.

[3] Z. Ibrahim, I. Ferrané, P. Joly. Temporal Relation Analysis in Audiovisual Documents for Complementary Descriptive Information. *In proc. of AMR 2005*, Glasgow, UK, July 2005.

[4] B. Moulin. Conceptual graph approach for the representation of temporal information in discourse. *Knowledge based systems*, vol. 5, n°3, pp 183–192, 1992.

[5] M. Vilain, and H. A. Kautz. Constraint propagation algorithms for temporal reasoning. *In AAAI-86*, pp. 132-144, 1986.

[6] Y. Zhai, Z. Rasheed, M. Shah. Conversation Detection in Feature Films Using Finite State Machines. *In 17th International Conference on Pattern Recognition ICPR'04*, vol. 4, pp. 458-461, 2004.

Session IX

Poster

Mardi 13 juin 2006 - 16h45 18h00

Généralisation du noyau GLDS pour la Vérification du locuteur par SVM

Jérôme Louradour (1), *Khalid Daoudi* (1) *et Francis Bach* (2)

(1) IRIT, CNRS UMR 5505, Université Paul Sabatier, Toulouse, France
(2) Centre de Morphologie Mathématique, Ecole des Mines de Paris, Fontainebleau, France
Mél : louradou,@irit.fr, daoudi@irit.fr, francis.bach@mines.org

ABSTRACT

The Generalized Linear discriminant Sequence (GLDS) kernel provides good performance in SVM speaker verification in NIST SRE (Speaker Recognition Evaluation) evaluations. It is based on an explicit mapping of each sequence to a single vector in a feature space using polynomial expansions. Because of practical limitations, these expansions has to be of degree less or equal to 3. In this paper, we generalize the GLDS kernel to allow not only any polynomial degree but also any expansion (possibly infinite dimensional) that defines a Mercer kernel (such as the RBF kernel). We conceive a new kernel, and makes it tractable using a method of data reduction adapted to kernel methods : the Incomplete Cholesky Decomposition (ICD). We present experiments on NIST SRE database, that show good perspective for our new approach.

1. INTRODUCTION

Les algorithmes de classification Machines à Vecteurs de Support (SVM) sont aujourd'hui considérés comme une des méthodes les plus performantes pour de nombreux problèmes réels de classification et de régression. A l'origine, ils ont été conçus pour construire une fonction discriminante permettant de séparer au mieux des régions complexes, dans des problèmes de classification binaire. En cela ils constituent une alternative intéressante aux approches génératives classiques MMG (Mélanges de Modèles Gaussiens) pour la vérification du locuteur. Alors que les modèles génératifs sont particulièrement adaptés lorsque les données à classifier sont des séquences de vecteurs, de longueurs variables, un des principaux défis pour appliquer une approche discriminante à la biométrie à partir d'un extrait de parole est de l'adapter au traitement de données séquentielles.

Une des manières intuitives d'appliquer un SVM à la vérification du locuteur serait d'utiliser la même démarche qu'avec les modèles génératifs, c'est-à-dire d'apprendre des modèles discriminants dans l'espace des vecteurs acoustiques et de combiner les scores en phase de test pour décider de la classe d'une séquence. Mais malgré les récents efforts pour faire correspondre les scores SVM à des probabilités a posteriori [10], permettant ainsi de les combiner de manière naturelle via le théorème de Bayes, un problème en reconnaissance du locuteur reste de pouvoir exploiter des corpus d'apprentissage importants afin de caractériser au mieux les classes (locuteur / reste du monde) fortement multimodales. En effet, la complexité des algorithmes d'apprentissage empêche les SVM de profiter d'une affluence de données de développement, à moins d'avoir recours à des méthodes de réduction de données. Les approches par clustering [7] ne donnent pour l'instant pas de bonnes performances pour la vérification en milieu bruité comme c'est le cas dans les évaluation NIST SRE. Une alternative à l'application hasardeuse de réduction de vecteurs est l'utilisation de noyaux de séquences, qui permettent d'exprimer le critère d'apprentissage de manière adéquate à notre problème de classification de séquences, et ainsi de compacter judicieusement l'information structurée disponible en phase de développement. Notons qu'en mode "indépendant du texte" (sans information a priori sur le contenu prononcé dans les extraits à classer), les modèles stationnaires (MMG), qui considèrent les observations indépendants les unes des autres, donnent des performances similaires aux modèles dynamiques comme les modèles de Markov. Pour cela, nous nous limitons pour la vérification du locuteur aux noyaux entre ensembles de vecteurs, invariants à l'ordre chronologique des vecteurs [1], et que nous appelons dans la suite "noyau de séquences" par simplicité.

L'utilisation de noyaux de séquences pour la vérification du locuteur a été l'objet de plusieurs recherches ces dernières années. Dans [6, 12] par exemple, des noyaux de séquences basés sur des modèles génératifs ont été utilisés. Ces noyaux restent d'une complexité élevée dans le cas des MMG, communément utilisés pour capturer la distribution des paramètres acoustiques. Un noyau efficace qui a montré des performances prometteuses aux évaluations NIST SRE est le noyau GLDS [3]. Il consiste simplement en une projection explicite des séquences dans un espace de dimension fixe en utilisant une expansion polynomiale, suivie d'un produit scalaire linéaire.

Le noyau GLDS a cependant une limitation pratique et théorique. La première est que l'utilisation d'expansions polynomiales au delà d'un degré 3 n'est pas envisageable pour des problèmes de grande dimensionnalité. La seconde vient du fait qu'il n'est pas possible de généraliser l'approche en l'état à des expansions infinies (l'astuce du noyau n'est pas vraiment utilisée). Le but de cet article est d'aller au-delà de ces deux limitations. Nous commencons par définir une classe de noyaux de séquences dont le GLDS est un cas particulier, et développons une forme à dimension finie. Cette forme ayant une complexité élevée ("intractable") pour une application de vérification du locuteur, nous réduisons cette complexité grâce à une méthode de décomposition matricielle : la factorisation de Cholesky incomplète.

[1] Notons tout de même que l'information dynamique à court-terme est prise en compte dans les dérivées incluses dans ces vecteurs

2. GÉNÉRALISATION DU NOYAU GLDS

2.1. Présentation du noyau GLDS

La forme originale du noyau GLDS [3] fait intervenir une expansion polynomiale ϕ_p, composée de monômes jusqu'à un degré donné p. Par exemple, si $p = 2$ et $\mathrm{x} = [x_1, x_2]^\top$ est un vecteur à deux dimensions, $\phi_p(x) = [x_1, x_2, x_1^2, x_1x_2, x_2^2]^\top$. Le noyau entre deux séquences de vecteurs $X = \{\mathrm{x}_t\}_{t=1...T_X}$ et $Y = \{\mathrm{y}_t\}_{t=1...T_Y}$ est donné comme un produit scalaire normalisé entre expansions moyennes :

$$K(X,Y) = \tfrac{1}{T_X} \sum_{t=1}^{T_X} \phi_p(\mathrm{x}_t)^\top \ \mathbf{M}_p^{-1} \ \tfrac{1}{T_Y} \sum_{s=1}^{T_Y} \phi_p(\mathrm{y}_s) \quad (1)$$

où $\mathbf{M}_p$ est la matrice moment d'ordre 2 des expansions polynomiales ϕ_p estimée sur une population de développement, ou son approximation diagonale pour plus d'efficacité.

Conçu de cette manière, le noyau GLDS n'offre pas beaucoup de flexibilité pour coller aux données (peu de paramètres réglables), et l'expansion ϕ_p atteind des dimensions trop élevées pour un degré p supérieur à 3. Un problème intéressant est de trouver une manière d'implémenter (1) pour n'importe quel p, quitte à faire des approximations. Un problème plus général est d'aboutir à une forme finie de (1) pour n'importe quelle expansion ϕ, y compris les expansions infinies, de manière à pouvoir supporter des noyaux de types RBF. C'est le but de la prochaine section.

2.2. Une classe riche de noyaux

Nous considérons la classe de noyaux de séquences de la forme :

$$\begin{aligned} \hat{K}(X,Y) &= \tfrac{1}{T_X T_Y} \sum_{t=1}^{T_X} \sum_{s=1}^{T_Y} \phi(x_t)^\top \mathbf{M}^{-1} \phi(y_s) \\ &= \overline{\phi}(X)^\top \mathbf{M}^{-1} \overline{\phi}(Y) \end{aligned} \quad (2)$$

où :

- ϕ est une expansion vectorielle de taille $D \leq +\infty$ définissant un noyau de Mercer k :

$$k(\mathrm{x},\mathrm{y}) = \phi(\mathrm{x})^\top \phi(\mathrm{y}) \quad (3)$$

- $\overline{\phi}$ désigne la moyenne sur une séquence des expansions vectorielles ϕ (même notation dans la suite).
- $\mathbf{M} = E(\phi\phi^\top)$ est la matrice des moments d'ordre 2 des expansions ϕ estimée sur une population $B = \{\mathrm{b}_1, \ldots, \mathrm{b}_n\}$ de taille n. $\mathbf{M}$ peut être exprimée comme un produit matriciel faisant intervenir la matrice $(D \times n)$ des expansions de B, $\Phi_B = [\phi(\mathrm{b}_1), \ldots, \phi(\mathrm{b}_n)]$:

$$\mathbf{M} = \frac{1}{n} \Phi_B \Phi_B^\top \quad (4)$$

Remarquons que $\hat{k}(\mathrm{x},\mathrm{y}) = \phi(\mathrm{x})^\top \mathbf{M}^{-1} \phi(\mathrm{y})$ définit aussi un noyau satisfaisant la condition de Mercer, avec une normalisation dans l'espace caractéristique (*feature space*) défini par l'expansion ϕ. Le noyau de séquences peut être réécrit comme somme linéaire de ce noyau repondéré : $\hat{K}(X,Y) = \frac{1}{T_X T_Y} \sum_t \sum_s \hat{k}(\mathrm{x}_t, \mathrm{y}_s)$.

2.3. Expression de $\hat{K}$ dans la forme duale

Dans cette section, nous montrons comment exprimer le noyau vectoriel normalisé $\hat{k}$ en fonction du noyau vectoriel standard k défini en (3) et des données de normalisation $B = \{\mathrm{b}_1, \ldots, \mathrm{b}_n\}$.

Considérons la Décomposition en Valeur Singulière (SVD) mince [5] de la matrice des expansions Φ_B :

$$\Phi_B = USV^\top \quad (5)$$

où U and V sont des matrices orthogonales de tailles respectives $D \times r$ and $n \times r$, $r \leq \min(n, D)$ étant le rang de Φ_B. Nous pouvons décomposer de cette manière :

$$\mathbf{M} = \frac{1}{n} USV^\top \ VSU^\top = \frac{1}{n} US^2U^\top \quad (6)$$

Remarquons que dans le cas général, $\mathbf{M}$ n'est pas nécessairement inversible et doit être régularisée, en utilisant par exemple $\mathbf{M} = E(\phi\phi^\top) + \frac{1}{n}\varepsilon I$. Cette régularisation s'impose pour des raisons statistiques dans les cas où la dimension D est plus grande que le nombre de vecteurs n [11]. Toutefois, les approximations faites en section 3 permettent de se passer de cette régularisation. Nous nous limitons donc ici à une pseudo-inversion de (6) :

$$\mathbf{M}^{-1} = nUS^{-2}U^\top = n\Phi_B V S^{-4} V^\top \Phi_B^\top \quad (7)$$

La matrice de *Gram* sur B, définie par $\mathbf{K}_{i,j} = k(\mathrm{b}_i, \mathrm{b}_j)$, peut s'écrire $\mathbf{K} = \Phi_B^\top \Phi_B$. Selon (5), elle a une Décomposition en Valeur Singulière explicite $\mathbf{K} = VS^2V^\top$. En considérant la pseudo-inverse $\mathbf{K}^{-2} = VS^{-4}V^\top$, le noyau $\hat{k}$ peut s'écrire :

$$\begin{aligned} \hat{k}(x,y) &= n \ \ \phi(x)^\top \Phi_B \ \ \mathbf{K}^{-2} \ \ \Phi_B^\top \phi(y) \\ &= n \ \ \Psi_B(x)^\top \ \ \mathbf{K}^{-2} \ \ \Psi_B(y) \end{aligned} \quad (8)$$

où l'on définit la projection vectorielle de taille n via le noyau (3) :

$$\Psi_B(\mathrm{x}) = [k(\mathrm{b}_1, \mathrm{x}), \ldots, k(\mathrm{b}_n, \mathrm{x})]^\top \quad (9)$$

Par linéarité dans l'espace caractéristique, nous pouvons finalement écrire le noyau de séquences $\hat{K}$ dans une forme finie :

$$\hat{K}(X,Y) = n \ \overline{\Psi}_B(X)^\top \ \mathbf{K}^{-2} \ \overline{\Psi}_B(Y) \quad (10)$$

En pratique, le nombre de vecteurs disponibles pour le traitement de parole peut être très grand, et la taille n peut être énorme. L'implémentation de $\hat{K}$ en utilisant (10), avec une complexité en $O\left(n^2\right)$, peut donc vite être intractable. Dans la section suivante, nous utilisons une décomposition matricielle pour remédier à ce problème en donnant une forme approchée mais traitable de (10).

3. RÉDUCTION DE DIMENSIONNALITÉ

Les méthodes de réductions de données pour les méthodes à noyaux correspondent à des approximations de la matrice de Gram [4]. Le but de ces méthodes est de choisir un sous-ensemble $C \subset B$ qui permettrait d'approcher la matrice de Gram avec un rang inférieur, de manière à reformuler le problème avec une complexité moindre. Parmi ces techniques, la Factorisation de Cholesky Incomplète (ICD) [1] a une complexité relativement basse, en $O(m^2 n)$ si m est la taille désirée pour le sous-ensemble C. De plus, elle ne requiert pas le stockage en mémoire de l'intégralité de la matrice $\mathbf{K}$.

Etant donnée une matrice de Gram $\mathbf{K}$ de taille $n \times n$ (le rang de $\mathbf{K}$ pouvant être plus faible que n), l'ICD de $\mathbf{K}$ est une matrice $\mathbf{G}_m$ de taille $n \times m$ (de rang $m < n$),

telle que $\mathbf{K}$ paut être approchée par $\mathbf{G}_m{\mathbf{G}_m}^\top$. La racine $\mathbf{G}_m$ est générée par les colonnes de $\mathbf{K}$ indexées par $I = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$. Ainsi, l'on peut considérer que l'ICD fournit un *codebook* $C = \{\mathrm{b}_{i_1}, \ldots, \mathrm{b}_{i_m}\} \subset B$. Dans la suite, nous montrons comment exprimer notre noyau de séquences par une forme traitable utilisant C au lieu de B. Il peut être montré [1] que $\mathbf{G}_m$ peut s'écrire :

$$\mathbf{G}_m = \mathbf{K}(:, I)\mathbf{K}(I, I)^{-1/2} \tag{11}$$

où $\mathbf{K}(:, I)$ désigne toutes les colonnes de $\mathbf{K}$ indexées par I. Avec la même notation, $\mathbf{K}(I, I)$ est une matrice de Gram $m \times m$ sur les entrées $\{\mathrm{b}_{i_1}, \ldots, \mathrm{b}_{i_m}\}$.

Le fait que Φ_B et ${\mathbf{G}_m}^\top$ soient considérés comme ayant le même carré ($\mathbf{K} = \Phi_B^\top \Phi_B \approx \mathbf{G}_m{\mathbf{G}_m}^\top$) implique qu'il existe une matrice orthogonale U telle que l'on peut considérer la décomposition incomplète, à la place de (5) : $\Phi_B = U{\mathbf{G}_m}^\top$. La matrice $\mathbf{M}$ peut ainsi être approchée par $\frac{1}{n}U{\mathbf{G}_m}^\top\mathbf{G}_m U^\top$, ce qui revient à régulariser $\mathbf{M}$. Cette décomposition permet d'inverser :

$$\begin{aligned} \mathbf{M}^{-1} = &\quad n\, U\, ({\mathbf{G}_m}^\top \mathbf{G}_m)^{-1}\, U^\top \\ = &\quad U\ \mathbf{K}(I,I)^{1/2}\, \mathbf{R}^{-1}\, \mathbf{K}(I,I)^{1/2}\ U^\top \end{aligned} \tag{12}$$

où nous définissons selon (11) la matrice $m \times m$ (nécessairement inversible après l'ICD) :

$$\mathbf{R} = \frac{1}{n}\mathbf{K}(:, I)^\top \mathbf{K}(:, I) \tag{13}$$

Nous pouvons aussi déduire de (11) que $\Phi_B^\top = \mathbf{G}_m U^\top = \mathbf{K}(:, I)\mathbf{K}(I, I)^{-1/2}U^\top$. Si nous supposons que les expansions $\phi(\mathrm{x})$ appartiennent au sous-espace affine généré par les expansions incluses dans Φ_B, alors on peut généraliser

$$\phi(\mathrm{x})^\top = \Psi_C(\mathrm{x})^\top \mathbf{K}(I, I)^{-1/2} U^\top \tag{14}$$

où $\Psi_C(\mathrm{x})$ est la projection réduite sur les vecteurs de C dans l'espace caractéristique : $\Psi_C(\mathrm{x}) = [k(\mathrm{b}_{i_1}, \mathrm{x}), \ldots, k(\mathrm{b}_{i_m}, \mathrm{x})]^\top$. La formulation de notre nouveau noyau de séquences est finalement obtenue en injectant les nouvelles expressions de $\mathbf{M}^{-1}$ et $\phi(X)$ dans (2) :

$$\hat{K}_{ICDS}(X, Y) = \overline{\Psi}_C(X)^\top\ \mathbf{R}^{-1}\ \overline{\Psi}_C(Y) \tag{15}$$

La complexité de $\hat{K}_{ICDS}(X, Y)$ est en $O(m^2)$. En pratique, la valeur de m peut être choisie très inférieure à n, ce qui fait aboutir à une implémentation efficace d'un noyau.

Il est intéressant de remarquer que le noyau de séquences $\hat{K}_{ICDS}$ donné par (15) a une forme similaire à notre noyau de séquences RKHS défini dans notre précédent travail [8], où nous avions adopté la même stratégie que Campbell dans [3] pour concevoir un noyau mesurant la similarité entre deux séquences. Cette stratégie consiste à apprendre un modèle discriminant (avec pour valeurs cibles 0/1) sur une séquence (dans un Espace de Hilbert à Noyaux Reproduisants généré par k), et à l'appliquer sur une autre en supposant l'indépendance des observations. Après quelques approximations, un noyau symétrique vérifiant les conditions de Mercer est obtenu.

4. RÉSULTATS EXPÉRIMENTAUX

4.1. Corpus et pré-traitement

Nous testons notre système sur les données de NIST SRE 2004, en utilisant le protocole de développement défini par le projet Biosecure [2]. Dans ce protocole, nous considérons 113 locuteurs imposteurs pour développer le système. L'évaluation comprend plus de 7000 tests impliquant 181 locuteurs cibles et 368 séquences de test.

Les vecteurs acoustiques extraits des séquences de parole sont 12 MFCC auxquels nous rajoutons les dérivées temporelles premières. Un extracteur de silence utilisant un modèle bigaussien non-supervisé permet de rejeter les vecteurs correspondant aux segments de silence. La méthode de normalisation utilisée, après suppression des silences, est le *feature warping* [9].

4.2. Description du système

Une fois un noyau k choisi, la première étape est d'appliquer l'ICD à la matrice de gram estimée sur les données de développement. Dans les résultats montrés dans cet article, nous sélectionnons aléatoirement $n = 20000$ vecteurs (parmi les plus de 200000 disponibles) sur lesquels la décomposition est appliquée ; Des expérimentations ont montré que les performances étaient peu sensibles au nombre n de données de développement considérées du moment qu'il est suffisamment grand par rapport au nombre m de vecteurs retenus en sortie de l'ICD. Pour un bon compromis entre complexité et performance, nous avons choisi de retenir $m \sim 5000$ vecteurs *codebook*. Une fois la projection $\overline{\Psi}_C$ déterminée par le choix de k et du *codebook* C, la matrice de normalisation $\mathbf{R}$ définie par (13) peut être calculée.

Les modèles SVM des locuteurs cibles sont entrainés en considérant un ensemble commun de séquences imposteurs, dont les caractéristiques (valeurs du noyau (15) entre toutes les paires de séquences) peuvent être calculées à l'avance et stockées en mémoire. Pour rendre plus efficace l'apprentissage des modèles, les projections des séquences imposteurs sont mémorisées sous la forme $\mathbf{R}^{-1}\overline{\Psi}_C$. De cette manière, quand une séquence S émise par un locuteur cible est donnée au système, il n'y a qu'à calculer $\overline{\Psi}_C(S)$ pour obtenir les valeurs du noyau avec les séquences imposteurs par un simple produit scalaire.

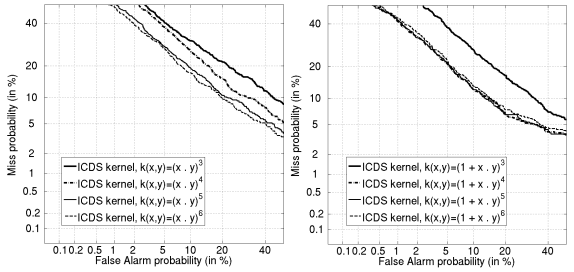
La procédure de test peut être rendue efficace par une astuce similaire. La fonction discriminante pour un locuteur peut être compactée dans un vecteur de dimension m (de manière analogue à [8]).

4.3. Choix des paramètres du noyau

Dans cette section, nous discutons comment choisir les paramètres du noyau k choisi pour définir $\hat{K}$.

Noyaux polynomiaux de la forme $k_p(\mathrm{x}, \mathrm{y}) = (c + \mathrm{x}\cdot\mathrm{y})^p$ Les résultats correspondants aux valeurs $c = 0$ (Fig.1.a) et $c = 1$ (Fig.1.b) montrent qu'il vaut mieux prendre un valeur non nulle pour c. Il est donc préférable de tenir compte de tous les monômes avec un degré inférieur ou égal à p, comme avec le noyau GLDS (quand $c = 0$, seulement les monômes de degrés p sont pris en compte).

De plus, les Figures 1.a et 1.b montrent que les performances sont meilleures avec un degré plus grand que 3, ce qui suggère que le noyau GLDS donnerait de meilleures performances en considérant aussi les monômes d'un degré supérieur à 3 dans l'expansion ϕ_p.



(a) forme $(x \cdot y)^p$ (b) forme $(1 + x \cdot y)^p$

FIG. 1: Performances des noyaux polynomiaux

Noyaux RBF de la forme $k_{rbf}(\mathrm{x}, \mathrm{y}) = e^{-\gamma|\mathrm{x}-\mathrm{y}|^2}$ [11] recommande de choisir γ de l'ordre de $\gamma_0 = 1/2d\sigma^2$, où σ est la moyenne des écarts-types pour les composantes des vecteurs d'observation. Avec notre pré-traitement, ceci correspond à $\gamma_0 \approx 0.3$. Nos expériences confortent ce choix (fig.2). En fait, si γ est trop élevé, le noyau vectoriel colle trop aux données, $\mathbf{K}$ est proche de la matrice identité (rang maximal), et la projection de séquence définie en (15) revient à compter combien de vecteurs de la séquence gisent dans le voisinage de chaque vecteur *codebook*. Au contraire, si γ est trop faible, le rang de $\mathbf{K}$ est faible, et un nombre trop faible de caractéristiques sera considéré.

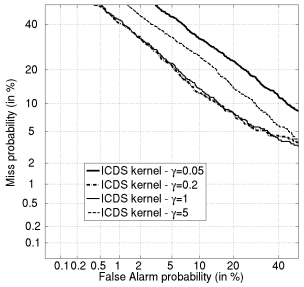


FIG. 2: Performances pour les noyaux RBF

4.4. Comparaison avec d'autres systèmes SVM

Les meilleures performances de notre noyau ICDS sont obtenus avec un noyau RBF (les résultats sont montrés avec $\gamma = 0.2$). Dans fig.3, elles sont comparées à celles d'autres noyaux de séquence (modélisation SVM) :

- L'approche GLDS [3] avec une expansion polynomiale de degré 3 et une approximation diagonale de la matrice second moment $\mathbf{M}_p$.
- La même approche avec une matrice pleine $\mathbf{M}_p$.
- Notre approche précédente [8] (noyau RKHSS).

Les résultats montrent que notre système donne de meilleures performances que les autres. Des expérimentations avec d'autre paramétrisation (exploitant l'information spectrale) confirment cette tendance.

5. CONCLUSION

Nous avons présenté une manière de généraliser le noyau GLDS à un espace caractéristique défini par un noyau vectoriel de Mercer quelconque. Le noyau de séquences ainsi conçu a été rendu implémentable en utilisant une décomposition de la matrice de gram correspondant au noyau vectoriel. Il conduit à des performances meilleures que le

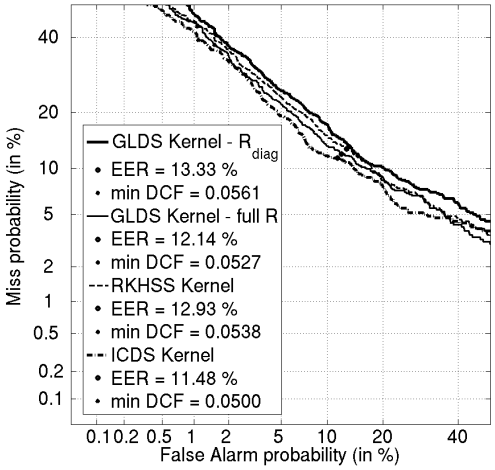


FIG. 3: Comparaison de plusieurs systèmes SVM

GLDS et que notre ancienne approche similaire. Plusieurs extensions sont possibles. Par exemple, il serait intéressant de considérer la matrice de covariance à la place de la matrice de second moment $\mathbf{M}$, afin de définir un noyau basé sur la distance de Malahanobis dans l'espace caractéristique.

RÉFÉRENCES

[1] F.R. Bach and M.I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proc. ICML*, 2005.

[2] Biosecure network of excellence : Biometrics for secure authentification. http ://www.biosecure.info, 2005.

[3] W.M. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 2005.

[4] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2, 2001.

[5] G.H. Golub and C.F. Van Loan. *Matrix Computation*. The John Hopkins Univ. Press, 1996.

[6] P. Ho and P Moreno. SVM kernel adaptation in speaker classification and verification. In *Proc. ICSLP*, 2004.

[7] Z. Lei, Y. Yang, and Z. Wu. Mixture of support vector machines for text-independent speaker recognition. In *Proc. Interspeech*, 2005.

[8] J. Louradour and K. Daoudi. Conceiving a new sequence kernel and applying it to SVM speaker verification . In *Proc. Interspeech*, 2005.

[9] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey*, 2001.

[10] J. Platt. *Probabilities for SV Machines*. MIT Press, 2000.

[11] B. Schölkopf, S. Mika, C. J.C.Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola. Input space versus feature space. *IEEE Trans. Neural Networks*, 100(5) :1000–1017, 1999.

[12] V. Wan. *Speaker Verification using Support Vector Machines*. PhD thesis, University of Sheffield, 2003.

REPRESENTATION DU LOCUTEUR PAR MODELES D'ANCRAGE POUR L'INDEXATION DE DOCUMENTS AUDIO

Mikaël Collet (1)(2), Delphine Charlet (1), Frédéric Bimbot (2)

(1) France Telecom R&D - TECH/SSTP - 2 av. Pierre Marzin - 22307 Lannion Cedex - FRANCE
{mikael.collet, delphine.charlet}@rd.francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
frederic.bimbot@irisa.fr

ABSTRACT

This paper presents a speaker indexing system of audio document entirely based on the anchor models approach. Evaluation is done on the audio database of the ESTER evaluation campaign for the rich transcription of French broadcast news. Results show that speaker indexing performance is improved when a speaker clustering process is performed and that a weighted measure of similarity, used in the speaker tracking process, can overcome some errors of the clustering process. The use of anchor models is particulary suitable for speaker indexing because the computational burden to search a speaker in an audio document is very low and performances are equivalent to those of a speaker indexing system using the classical speaker representation in the acoustic space (Gaussian model for speaker segmentation and clustering, Gaussian mixture model for speaker tracking).

1. INTRODUCTION

La représentation du locuteur par les modèles d'ancrage consiste à modéliser un locuteur relativement à un ensemble de locuteurs de référence appelés modèles d'ancrage. Ces modèles d'ancrage peuvent être vus comme une représentation du signal de parole (comme les coefficients cepstraux) particulièrement adaptée pour la caractérisation du locuteur. Cette représentation du locuteur est également appropriée pour la recherche de locuteurs cibles au sein de grandes bases de données audio [1]. En effet, dans le cadre de l'utilisation des modèles d'ancrage, la majorité des calculs est effectuée lors d'une phase de pré-traitement de la base de données (segmentation et regroupement en locuteurs), indépendamment des locuteurs à rechercher. Le coût de calcul pour rechercher des nouveaux locuteurs cibles est alors très faible.
Cet article, qui présente un système d'indexation en locuteurs de documents audio entièrement basé sur la représentation du locuteur par les modèles d'ancrage, est organisé de la manière suivante. Le paragraphe 2, présente le concept des modèles d'ancrage et les méthodes permettant de comparer des locuteurs dans l'espace des modèles d'ancrage. Ensuite les deux processus de pré-traitement de la base de données audio sont décrits : la segmentation en locuteurs (paragraphe 3) et le regroupement en locuteurs (paragraphe 4). Puis le paragraphe 5 présente différents processus de suivi de locuteurs détaillés dans des travaux récents [2] [3]. Enfin, au cours du paragraphe 6, les performances de ces processus de suivi de locuteurs sont évaluées sur la base de donnée d'émissions radiophoniques de la campagne d'évaluation ESTER.

2. REPRÉSENTATION DU LOCUTEUR PAR MODÈLES D'ANCRAGE

Le locuteur constitue une clé d'indexation très pertinente pour l'archivage de documents audio (archives radiophoniques, messageries vocales...). Cependant, les techniques actuelles pour la représentation du locuteur (modélisation par un mélange de gaussiennes - GMM) sont assez gourmandes en temps de calcul et en taille mémoire et ne sont pas adaptées à l'indexation de grandes bases de données audio.
C'est pourquoi cet article propose un système d'indexation en locuteurs utilisant les modèles d'ancrage, permettant ainsi de diminuer le nombre de paramètres pour la représentation du locuteur tout en conservant les performances d'une modélisation par mélange de gaussiennes.

2.1. Principe

Des recherches récentes [1] se sont orientées vers une représentation relative du locuteur. Cette modélisation consiste à projeter un énoncé d'un locuteur dans un espace de locuteurs de référence. Le locuteur n'est plus représenté de façon absolue mais relativement à un ensemble de locuteurs caractérisés par des modèles GMM. Ces modèles sont appelés modèles d'ancrage.
Le locuteur est caractérisé par un vecteur défini comme l'ensemble des rapports de vraisemblance des données du locuteur issues des différents modèles d'ancrage. Ce vecteur est appelé *Speaker Characterization Vector* (SCV) et dénoté $\widetilde{X}$.

$$\widetilde{X} = \begin{bmatrix} \widehat{s}(X|\overline{\lambda}_1) \\ \widehat{s}(X|\overline{\lambda}_2) \\ \vdots \\ \widehat{s}(X|\overline{\lambda}_E) \end{bmatrix} \quad (1)$$

où $\widehat{s}(X|\overline{\lambda}_e)$ est le logarithme du rapport de vraisemblance des données X (de N vecteurs acoustiques) pour le modèle GMM du locuteur de référence $\overline{\lambda}_e$ (modèles d'ancrage) relativement à un modèle indépendant du locuteur dit "modèle du monde" (ou UBM - Universal Background Model) :

$$\widehat{s}(X|\overline{\lambda}_e) = \frac{1}{N} \log \frac{p(X|\overline{\lambda}_e)}{p(X|\lambda_{UBM})} \quad (2)$$

2.2. Mesures de similarité dans l'espace des modèles d'ancrage

Les mesures de similarité précédemment utilisées pour comparer des SCV dans l'espace des modèles d'ancrage sont la mesure de similarité euclidienne et la mesure de similarité angulaire. Une nouvelle mesure de similarité basée sur le coefficient de corrélation est également proposée dans [4].
Soient X and Y deux segments de parole, $\widetilde{X}$ et $\widetilde{Y}$ leur SCV. La mesure de similarité de corrélation est définie par :

$$\rho(\widetilde{X},\widetilde{Y}) = 1 - R(x,y) \quad (3)$$

où $R(x,y)$ est le coefficient de corrélation entre les composantes des deux SCV, les composantes étant considérées comme la réalisation de deux variables aléatoires x et y : $R(x,y) = \frac{C_{xy}}{\sigma_x \sigma_y}$
Cette nouvelle mesure de similarité s'avère expérimentalement plus robuste à la variabilité intra-locuteurs que les mesures de similarité euclidienne et angulaire [4].

2.3. Approche statistique dans l'espace des modèles d'ancrage

Le principe de l'approche statistique proposée dans [5] consiste à modéliser les différents énoncés d'un même locuteur par une distribution normale dans l'espace des modèles d'ancrage. Il s'agit d'un modèle statistique modélisant la variabilité intra-locuteur. En pratique, l'approche proposée dans [5] consiste à représenter un locuteur X ayant prononcé un ou plusieurs énoncés par une distribution notée $\hat{X}$ dans l'espace des SCV :

$$\hat{X} = \mathcal{N}(\mu_X, \Sigma_X) \quad (4)$$

où $\mathcal{N}$ est une distribution gaussienne de vecteur moyen μ_X et de matrice de covariance Σ_X. La matrice de covariance est la même pour tous les locuteurs et est égale à Σ_0. Le vecteur moyen du modèle du locuteur est adapté à partir d'un vecteur moyen μ_0 par une version simplifié de l'adaptation par maximum a posteriori (MAP). Les paramètres de la distribution a priori (μ_0 et Σ_0) sont estimés à partir d'un corpus de développement selon le processus décrit dans [5]. En vérification du locuteur, le score d'un segment de test Y pour un locuteur X est un log-rapport de vraisemblance entre le modèle du locuteur X et le modèle a priori $\mathcal{N}(\mu_0, \Sigma_0)$.

$$LLR(\widetilde{Y}|\hat{X}) = \log \frac{p(\widetilde{Y}|\mu_X, \Sigma_0)}{p(\widetilde{Y}|\mu_0, \Sigma_0)} \quad (5)$$

Une extension de cette approche statistique, proposée dans [5], consiste à estimer un modèle $\hat{Y}$ à partir du segment de test Y et à utiliser les données d'apprentissage $\widetilde{X}$ pour symétriser la mesure :

$$L(\widetilde{Y},\widetilde{X}) = \frac{LLR(\widetilde{Y}|\hat{X}) + LLR(\widetilde{X}|\hat{Y})}{2} \quad (6)$$

3. SEGMENTATION EN LOCUTEURS

Le processus de segmentation en locuteur consiste à segmenter un document audio en segments homogènes de longueur raisonnable ayant été prononcés par un seul locuteur. Ce processus s'effectue sans aucune connaissance a priori sur les locuteurs présents dans le document. Le processus de segmentation en locuteurs utilisé dans cet article est basé sur une technique qui consiste à détecter des ruptures statistiques dans le signal audio, correspondant à des changements de locuteurs [6]. Dans notre cas, le processus de segmentation en locuteurs utilise la représentation du locuteur par les modèles d'ancrage et le rapport de vraisemblance généralisé calculé dans l'espace des coefficients acoustiques est remplacé par la mesure de similarité de corrélation définie par l'équation 3. Des expériences préliminaires ont montré que les mesures de similarité dans l'espace des modèles d'ancrage, notamment la mesure de corrélation, obtiennent de meilleures performances en segmentation en locuteurs que le log-rapport de vraisemblance dans l'espace des modèles d'ancrage (équation 6). Les travaux présentés dans [2] montrent également que la mesure de similarité de corrélation dans l'espace des modèles d'ancrage permet de mieux détecter les changement de locuteurs que le rapport de vraisemblance généralisé dans l'espace des coefficients acoustiques.

4. REGROUPEMENT EN LOCUTEURS

Le processus de regroupement en locuteurs consiste à regrouper au sein d'une même classe les segments supposés prononcés par le même locuteur. Ce processus s'effectue sans aucune connaissance a priori sur les locuteurs présents dans le document. L'algorithme de regroupement en locuteurs présenté dans cet article est basé sur une approche par *single-linkage* [7] et s'effectue en trois étapes :

1. Calcul d'une mesure de similarité de corrélation (équation 3) entre tous les segments issus du processus de segmentation en locuteurs.
2. Regroupement dans une même classe d'un segment et de son plus proche voisin au sens de la mesure de similarité.
3. Fusion des classes dont l'intersection n'est pas vide.

5. SUIVI DE LOCUTEURS

Le processus de suivi de locuteurs au sein d'un document audio consiste à rechercher les intervalles de temps au cours desquels un locuteur cible a parlé. La majorité des systèmes présentés dans la littérature utilisent un processus de vérification du locuteur où les énoncés de test sont les segments issus d'un module de segmentation en locuteurs. La représentation du locuteur par mélange de gaussiennes (GMM) dans l'espace des coefficients acoustiques [8] constitue l'état de l'art pour les processus de suivi de locuteurs. Cependant, le coût de calcul de recherche d'un locuteur cible utilisant l'approche GMM, équivalent au calcul d'une vraisemblance dans un espace multi-gaussien pour chaque trame acoustique du document audio, est très important. Ainsi, afin de réduire le coût de calcul, on utilise l'approche statistique dans l'espace des modèles d'ancrage qui obtient des performances en vérification du locuteur équivalente à l'approche GMM [5]. Cette approche est utilisée dans les trois processus de suivi de locuteurs présentés au cours des paragraphes suivants. Dans ce cas, le coût de calcul est équivalent au calcul d'une vraisemblance dans un espace mono-gaussien pour chaque segment du document audio. Dans la suite de cet article, nous considérons les notations suivantes : soit X un locuteur cible, $Y_i, i = 1, \ldots, N$ les segments issus du module de segmentation et C_{Y_i} une classe de $N_{C_{Y_i}}$ segments à laquelle appartient le segment Y_i.

5.1. Suivi sans regroupement en locuteurs

Le processus de suivi de locuteurs sans regroupement en locuteurs consiste à comparer les segments Y_i au locuteurs X en utilisant le log-rapport de vraisemblance défini par l'équation 6.

5.2. Suivi avec regroupement en locuteurs

Le processus de suivi de locuteurs avec regroupement en locuteurs utilise pour chaque segment issu du processus de segmentation en locuteurs une information d'appartenance à une classe de locuteurs. L'information d'appartenance à une classe de locuteurs est déterminée par le processus de regroupement en locuteurs présenté au paragraphe 4. Cette méthode, présentée dans [2], permet d'obtenir des mesures de similarité plus fiables, notamment pour les segments de parole courts. La mesure de similarité entre le locuteur cible X et un segment Y_i en fonction de la classe C_{Y_i} est définie par :

$$L_{C_{Y_i}}(\widetilde{X},\widetilde{Y}_i) = \frac{1}{N_{C_{Y_i}}} \sum_{j=1}^{N_{C_{Y_i}}} L(\widetilde{X},\widetilde{Y}_j) \qquad (7)$$

où les segments Y_j appartiennent à la classe C_{Y_i}.

5.3. Suivi avec regroupement en locuteurs et mesure de similarité pondérée

Des expériences présentées dans [3] ont montré que les erreurs du processus de regroupement en locuteurs engendrent de nouvelles erreurs de suivi de locuteurs. Une nouvelle mesure de similarité dite *pondérée* permet de limiter la contribution des segments Y_j dans le calcul de l'équation 7 lorsque la probabilité que Y_j ait été prononcé par le même locuteur que Y_i est faible. La réduction des erreurs de suivi dues aux erreurs de regroupement s'effectue en introduisant un coefficient de pondération basé sur une mesure de similarité entre les segments Y_i et Y_j, relié à la probabilité que les segments soient prononcés par un même locuteur :

$$L^p_{C_{Y_i}}(\widetilde{X},\widetilde{Y}_i) = \frac{1}{\sum_{j=1}^{N_{C_{Y_i}}} \gamma_{ij}} \sum_{j=1}^{N_{C_{Y_i}}} \gamma_{ij} L(\widetilde{X},\widetilde{Y}_j) \qquad (8)$$

En pratique, le coefficient de pondération γ_{ij} est défini comme la fonction d'erreur complémentaire de la mesure de similarité de corrélation $\rho(\widetilde{Y}_i,\widetilde{Y}_j)$.

6. EXPERIENCES ET RESULTATS

Le système d'indexation en locuteurs utilisant la représentation du locuteur par les modèles d'ancrage est évalué sur la tâche de suivi de locuteurs de la campagne d'évaluation de systèmes de transcription d'émissions radiophoniques (campagne ESTER [9]). Le corpus d'évaluation, les mesures d'évaluation, la configuration et les performances du système sont présentés au cours des paragraphes suivants.

6.1. Corpus d'évaluation

Le corpus utilisé pour ces expériences est un corpus d'émissions radiophoniques en français. Le corpus est divisé en un ensemble d'apprentissage, un ensemble de développement et un ensemble de test selon les spécifications de la campagne ESTER (voir [9] pour plus de détails). L'ensemble d'apprentissage contient 82h d'émissions radiophoniques enregistrées sur la période 1998-2003. L'ensemble de développement contient 10h d'émissions radiophoniques enregistrées en 2003 et l'ensemble de test contient 10h d'émissions radiophoniques enregistrées en 2004. Les expériences présentées dans cet article sont effectuées sur l'ensemble de développement avec une liste de 279 locuteurs cibles fourni par les organisateurs de la campagne ESTER.

6.2. Mesure d'évaluation

6.2.1. Segmentation et regroupement en locuteurs

Les performances des systèmes de segmentation et de regroupement en locuteurs sont évaluées en termes de pureté moyenne en locuteurs P_{Loc} et de pureté moyenne des classes P_{Cl} . Ces mesures d'évaluations proposées par [10] sont définies par les équations suivantes :

$$P_{Cl} = \frac{1}{N_0} \sum_{i=1}^{N} p_i^{Cl} n_i \text{ avec } p_i^{Cl} = \sum_{j=1}^{S} \frac{n_{ij}^2}{n_i^2} \qquad (9)$$

$$P_{Loc} = \frac{1}{N_0} \sum_{j=1}^{S} p_j^{Loc} n_j \text{ avec } p_j^{Loc} = \sum_{i=1}^{N} \frac{n_{ij}^2}{n_j^2} \qquad (10)$$

où N est le nombre de classes du document audio, S le nombre de locuteurs du document audio, N_0 le nombre de trames du document audio, n_i le nombre de trames de la classe i, n_j le nombre de trames du locuteur j, n_{ij} le nombre de trames dans la classe i prononcées par le locuteur j.

6.2.2. Suivi de locuteurs

Les performances du système de suivi de locuteurs sont évaluées en termes de précision (P) et de rappel (R) :

- $P = \frac{\text{Nombre de trames du locuteur cible détectées}}{\text{Nombre de trames détectées}}$
- $R = \frac{\text{Nombre de trames du locuteur cible détectées}}{\text{Nombres de trames du locuteur cible}}$

Les valeurs de Précision et de Rappel sont combinées en une seule valeur d'évaluation en utilisant la F-mesure, qui est définie par

$$F = \frac{2.P.R}{P+R} \qquad (11)$$

6.3. Configuration du système

Dans chacune des expériences, 13 MFCC avec leurs dérivées premières et secondes plus ΔE and $\Delta\Delta E$ sont utilisés. Les modèles d'ancrage sont des modèles statistiques GMM à 256 gaussiennes appris par adaptation MAP d'un modèle UBM indépendent du genre. L'espace des modèles d'ancrage est composé de tous les locuteurs, différents des locuteurs cibles, qui ont plus de 70 secondes de parole disponibles dans l'ensemble d'apprentissage, soit 316 locuteurs. Aucune compensation de canal n'est appliquée pour les processus de segmentation et de regroupement en locuteurs et un module de *feature warping* [11] est appliqué sur les données acoustiques pour le processus de suivi de locuteurs.

6.4. Résultats

6.4.1. Segmentation et regroupement en locuteurs

Le tableau 1 présente les performances des processus de segmentation et de regroupement en locuteurs en termes de pureté moyenne des classes et pureté moyenne en locuteurs. Les erreurs de regroupement en locuteurs font diminuer la pureté des classes, cependant la pureté en locuteur est considérablement augmentée.

	P_{Cl}	P_{Loc}
Segmentation en locuteurs	95.3	19.2
Regroupement en locuteurs	83.7	54.3

Table 1. Performances des processus de segmentation et regroupement en locuteurs en termes de pureté moyenne en locuteurs et pureté moyenne des classes

6.4.2. Suivi de locuteurs

Trois systèmes de suivi de locuteurs sont comparés et leurs performances sont représentées sur la figure 1 en termes de précision et de rappel : suivi sans regroupement en locuteurs, suivi avec regroupement en locuteurs, suivi avec regroupement en locuteurs et mesure de similarité pondérée. Cette figure montre que le regroupement en locuteurs permet d'augmenter significativement les performances du système de suivi de locuteur pour atteindre des performances équivalentes à celles obtenues par un système de suivi de locuteurs utilisant une modélisation du locuteur par mélange de gaussiennes dans l'espace acoustique [12].
Le tableau 2, qui indique les points de fonctionnement optimaux de chaque système (point de la courbe précision/rappel qui maximise la F-mesure), montre que l'amélioration des performances liée au regroupement en locuteurs se traduit par une augmentation du taux de rappel (79.5 % contre 67.5 %) tandis que la mesure de similarité pondérée améliore la précision en réduisant l'impact des erreurs du processus de regroupement (86.8 % contre 84.4 %).

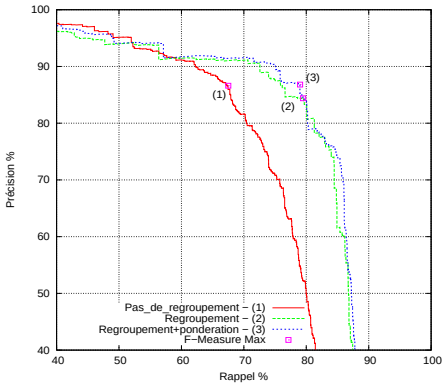


Fig. 1. Precision/Rappel pour chaque système de suivi de locuteurs : sans regroupement, avec regroupement, avec regroupement et mesure de similarité pondérée

Système	F_{max}	Précision	Rappel
Pas de regroupement (1)	75.8	86.6	67.5
Regroupement (2)	81.9	84.4	79.5
Regroupement + pondération (3)	82.7	86.8	79.0

Table 2. Points de fonctionnement des systèmes de suivi de locuteurs

7. CONCLUSION

Dans cet article, nous avons présenté un système d'indexation en locuteurs entièrement basé sur la représentation du locuteur par modèles d'ancrage qui consiste à représenter un locuteur dans un espace de locuteurs de référence. L'espace des locuteurs de référence est muni de méthodes permettant de comparer deux locuteurs utilisées dans des contextes différents : la mesure de similarité de corrélation [4] est utilisée pour les processus de segmentation et de regroupement en locuteurs tandis que l'approche statistique [5] est utilisée pour le processus de suivi de locuteurs.
Les évaluations montrent que l'utilisation d'un processus de regroupement pour le suivi de locuteurs améliore significativement le taux de rappel du système et que la mesure de similarité pondérée permet d'améliorer la précision en réduisant l'impact des erreurs de regroupement. Les performances du système d'indexation en locuteurs sont équivalentes aux performances d'un système d'indexation utilisant une représentation classique du locuteur dans l'espace des coefficients acoustiques [12]. Ce résultat est particulièrement intéressant car il montre que tout en conservant des performances équivalente à l'état de l'art [9], la représentation du locuteur par les modèles d'ancrage permet de diminuer le coût de calcul pour l'indexation de grandes bases de donnée audio.

8. REFERENCES

[1] D.E. Sturim, D.A. Reynolds, E. Singer, and J.P. Campbell, "Speaker indexing in large audio databases using anchor models," in *ICASSP2001*, 2001, pp. 429–432.

[2] M. Collet, D. Charlet, and F. Bimbot, "Speaker tracking by anchor models using speaker segment cluster information," in *ICASSP*, 2006.

[3] M. Collet, D. Charlet, and F. Bimbot, "A weighted measure of similarity for speaker tracking," in *Speaker Odyssey*, 2006.

[4] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models," in *ICASSP*, 2005.

[5] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic anchor models approach for speaker verification," in *INTERSPEECH*, 2005.

[6] P. Delacourt, D. Kryze, and C. Wellekens, "Speaker-based segmentation for audio data indexing," in *ESCA ETRW Workshop*, 1999.

[7] L. Couvreur and J. M. Boite, "Speaker tracking in broadcast audio material in the framework of the thisl project," in *ESCA ETRW*, 1999.

[8] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier, "The ESTER phase 2 evaluation campaign for the rich transcription of french broadcast news," *EUROSPEECH*, 2005.

[10] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *ICASSP*, 1998.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Speaker Odyssey*, 2001.

[12] D. Moraru, M. Ben, and G. Gravier, "Experiments on tracking and segmentation in radio broadcast news," in *INTERSPEECH*, 2005.

Application d'un algorithme génétique à la synthèse d'un prétraitement non linéaire pour la segmentation et le regroupement du locuteur

Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, Jean- Luc Zarader

Groupe Perception et Réseaux Connexionnistes (PRC)

Université Pierre et Marie Curie
Christophe.Charbuillet@lis.jussieu.fr, Bruno.Gas@upmc.fr, Mohamed.Chetouani@upmc.fr, zarader@ccr.jussieu.fr

ABSTRACT

Speech feature extraction plays a major role in a speaker recognition system. B. Gas & al. showed in [1] that a non linear filtering of speech can improve the feature extractor's ability. In this article we propose to use genetic algorithms to design a non-linear pre-processing of speech adapted to the speaker diarization task. The pre-processing system we present is based on artificial recurrent neural networks (ARNN). We used a genetic algorithm to find both the structure and the weights of the network.

Experiments, carried out using a state-of-the-art speaker diarization system, showed that the proposed method gives promising results.

1. INTRODUCTION

L'étape d'extraction de caractéristiques occupe une place fondamentale dans un système de reconnaissance du locuteur. Les méthodes d'analyse du signal de la parole couramment utilisées aujourd'hui se divisent en deux groupes: les méthodes basées sur une modélisation de la production de la parole (LPC, LPCC) ainsi que les méthodes modélisant le système auditif humain (PLP, MFCC).
Depuis quelques années, un certain nombre de travaux se sont portés sur l'application de techniques non linéaires au problème de l'extraction de caractéristiques. Pitsikalis & Maragos [2] ainsi que Lindgren & al. [3] ont mis en évidence que les outils issus de la théorie du chaos sont applicables pour modéliser les phénomènes dynamiques non linéaires présents dans le signal de la parole. D'autre part, M. Chetouani & B. Gas [1] ont mis eux aussi en évidence qu'une transformation non linéaire du signal de parole permettait de faciliter l'extraction de caractéristiques discriminantes pour la reconnaissance de la parole.

Notre étude s'inscrit particulièrement dans la continuité de cette dernière approche. La transformation proposée par B. Gas & al. repose sur l'utilisation d'un réseau de neurones de type MLP (Multi Layer Perceptron) qui constitue un filtre non linéaire à réponse impulsionnelle finie. Nous proposons dans le présent article d'étendre ce principe à un filtrage non linéaire à réponse impulsionnelle infinie par l'utilisation d'un réseau de neurones récurrent. L'objectif étant d'élaborer un prétraitement adapté à la tâche de segmentation et de regroupement du locuteur.

Les algorithmes génétiques (AG) ont été proposés par Holland en 1975 et sont aujourd'hui couramment utilisés dans divers domaines pour l'optimisation de systèmes complexes. L'application de cette famille d'algorithmes au domaine du traitement automatique de la parole a connu ces dernières années un succès grandissant. On pourra citer les travaux de Chin-Teng Lin & al. [4] portants sur l'application des AG au problème de la transformation de caractéristiques pour la reconnaissance de la parole, ainsi que ceux de Demirekler & Haydar [5] qui proposent d'utiliser ces algorithmes pour la sélection de caractéristiques destinées à la reconnaissance du locuteur.
Notre étude se base sur la capacité des AG à optimiser un système de façon non supervisé, sans connaissance à priori de son fonctionnement. L'algorithme possède donc une certaine autonomie dans ses moyens de résoudre le problème d'optimisation. Notre approche consiste à utiliser cette autonomie comme un outil d'exploration. Dans une précédente étude [6], cette méthodologie nous a permis de mettre en évidence l'importance de certaines informations spectrales pour la tache de segmentation et de regroupement du locuteur.

Nous présentons ici l'application d'un AG à la synthèse d'un prétraitement du signal de parole basé sur un réseau de neurones récurrent. L'objectif étant d'explorer les potentialités de ce type de traitement pour la tâche de segmentation et regroupement du locuteur.

Nous abordons dans la section 2 une description de la tâche de segmentation et regroupement du locuteur ainsi que le système mis en œuvre. Dans la section 3 nous décrivons le filtre de prétraitement. La 4[ème] section présente l'algorithme génétique utilisé. Les résultats obtenus sont exposés dans la section 5.

2. SEGMENTATION ET REGROUPEMENT DU LOCUTEUR

La tâche de segmentation et de regroupement du locuteur (SRL) consiste à identifier les segments de signal produits par le même locuteur. La phase de segmentation a pour objectif de détecter les moments de changement de locuteur. Elle est suivie d'une étape de regroupement qui consiste à étiqueter les segments obtenus en fonction du locuteur. Les applications de la SRL sont essentiellement tournées vers l'indexation de documents sonores.

Le système de segmentation et de regroupement mis en œuvre est basé sur le critère BIC (Bayesian Information Criterion), appliqué à une modélisation mono Gaussienne diagonale des vecteurs codes. Ces algorithmes sont implémentés dans l'outil logiciel audioseg [7]. Le principe du système de segmentation est de calculer la différence BIC entre deux fenêtres adjacentes du signal. Un changement de locuteur sera détecté à l'interface des fenêtres si la différence BIC est supérieure à zéro. Le fonctionnement de l'algorithme de regroupement est comparable au système de segmentation. Les segments ayant les différences BIC les plus faibles sont regroupés itérativement. L'algorithme s'arrête lorsque la distance minimale entre deux segments est supérieure à zéro. La différence BIC est donnés par:

$$dBIC(C_i, C_j) = -D_r(C_i, C_j) + \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log\left(\mathrm{n_i} + \mathrm{n_j}\right)$$

avec:

$$D_r(C_i, C_j) = \frac{n_i + n_j}{2}.\log\left|\Sigma_{ij}\right| - \frac{n_i}{2}.\log\left|\Sigma_i\right| - \frac{n_j}{2}.\log\left|\Sigma_j\right|$$

Où C_i, C_j sont deux séquences de vecteurs codes représentant deux segments; n_i est la dimension de C_i et Σ_i sa matrice de covariance. λ est un paramètre pénalisant la complexité du modèle évalué.

3. RESEAUX DE NEURONES RECURRENTS & PRETRAITEMENT NON LINEAIRE

Les réseaux de neurones récurrents (RNR) sont des systèmes dynamiques non linéaires. La richesse dynamique de ces systèmes suscite un fort intérêt dans de nombreuses disciplines. Les applications dans le domaine du traitement du signal sont nombreuses. On pourra citer les travaux de B. Cessac & al. [8] qui ont proposé une méthode exploitant les propriétés dynamiques d'un RNR en mode chaotique pour transmettre un signal entre deux neurones distants, ou encore ceux de H. Jaeger [9] qui ont mis en évidence l'apport des RNR au problème de la prédiction de systèmes non linéaires. Nous proposons dans cet article, d'utiliser les RNR comme filtre de prétraitement du signal de parole.

Le modèle du filtre mis en œuvre est défini par:

$$u_i(t+1) = \tanh\left(\sum_{j=1}^{N} \mathbf{J}_{ij}.u_j(t) + \mathbf{E}_j.e(t) + \mathbf{B}_j\right)$$

$$s(t) = u_N(t)$$

où e et s représentent respectivement le signal d'entrée et de sortie du filtre, N le nombre de neurones, $u_i(t)$ représente la sortie du neurone i au temps t, $\mathbf{J}_{ij}$ la connexion orientée du neurone j vers le neurone i, $\mathbf{E}_j$ la connexion de l'entrée vers le neurone j, $\mathbf{B}j$ la valeur du biais du neurone j.

4. ALGORITHME GENETIQUE

Un algorithme génétique est un outil d'optimisation. Son emploi permet de trouver les valeurs optimales d'un jeu de paramètres maximisant les performances du système. Les AG sont basés sur les théories Darwiniennes de l'évolution plus connues sous le nom de théorie de la sélection naturelle. Un AG opère sur une population d'individus. Dans notre application, les individus sont des filtres à RNR définis par leurs paramètres **J, E** et **B**. L'algorithme mis en œuvre est constitué de trois opérateurs: Sélection, Evaluation et Variation (SEV) [10]. Ces opérateurs sont appliqués à la population courante *p(t)*, produisant une nouvelle génération *p(t+1)*, par la relation *p(t+1)=SEV (p(t))*.

La première étape de l'algorithme consiste à initialiser aléatoirement les paramètres de chaque filtre de la population *p(0)*. Les opérateurs Sélection, Evaluation et Variation sont ensuite appliqués itérativement.

L'opérateur *Variation* consiste à faire varier les paramètres des filtres de la population. Dans notre application, un opérateur de création ou d'élimination de connexion est appliqué avec une probabilité P_{connex}, suivie par un opérateur de modification des poids des connexions, modifiant le poids d'une connexion ρ par**:** $\rho \leftarrow \rho + \alpha.\phi$ où α représente le facteur de variation du poids et ϕ est une variable aléatoire de distribution uniforme sur [-1 ; 1].

L'opérateur *Evaluation* est destiné à évaluer la performance de chaque individu de la population. Dans notre application, la performance sera quantifiée par le taux d'erreur de regroupement du locuteur du système intégrant le filtre de prétraitement à évaluer.

L'opérateur *Sélection* a pour fonction de sélectionner les *Ns* meilleurs individus de la population en fonction de leur performance. Ces individus seront ensuite dupliqués pour former une nouvelle population *p(t+1)* de *Np* individus.

L'application itérative de ces trois opérateurs aura pour effet d'améliorer la performance moyenne de la population dans le temps, et donc de produire des filtres de prétraitement adaptés au problème du

regroupement du locuteur. La figure n°1 illustre l'application de l'algorithme mis en oeuvre à l'optimisation du filtre de prétraitement.

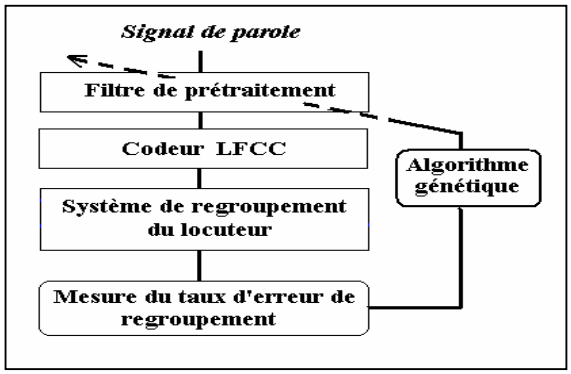


Figure 1 : Optimisation du filtre de prétraitement par algorithme génétique.

5. EXPERIENCES ET RESULTATS

Les expérimentations ont été menées sur la tache de segmentation et de regroupement du locuteur (SRL) de la campagne d'évaluation ESTER [11].

5.1. Bases de données

La base de donnée ESTER est composée de 40 heures de signaux audio provenant de l'enregistrement de 4 radios différentes échantillonnés à 16 kHz. Cette base est très riche: elle contient des signaux de parole enregistrés en studio, des interventions téléphoniques, de la parole sur font musical, avec des qualités d'enregistrement variables. Les fichiers audio ont d'une durée comprise entre 15 et 60 minutes. Le nombre de locuteur intervenant dans un fichier est compris entre 1 et 39.

La métrique de mesure des performances SRL est celle utilisée pour la campagne Nist RT 2003. Cette mesure est basée sur la qualité d'appariement du regroupement des locuteurs.

5.2. Bases d'évolution

Deux bases distinctes sont utilisées pour l'évolution des filtres. La première nommée "*base d'évolution"*, composée de 4 heures de signaux, est destinée à l'évaluation et à la sélection des filtres. La seconde nommée "*bases de cross validation",* composée de 8 heures d'enregistrement, a pour fonction de mesurer la capacité de généralisation des filtres obtenus.

5.3. Protocole expérimental

Les expériences d'évolution du filtre de prétraitement ont été menées sur la tâche de regroupement seul. La tâche de segmentation ayant été effectuée préalablement à partir d'un codage LFCC à 24 filtres et 16 coefficients) sans prétraitement. Le système de regroupement mis en œuvre comprend le filtre de prétraitement élaboré, suivi d'un codage LFCC 24/16 ainsi que du module de regroupement du locuteur. Les paramètres de l'algorithme génétique utilisés pour cette expérience sont: Np = 50; Ns = 15; N = 4; $\alpha = 0.05$; $P_{connex} = 0.01$;

L'initialisation de la population *p(0)* consiste à connecter toutes les connexions des réseaux et à initialiser leurs poids aléatoirement entre -0.6 et +0.6 pour les matrices **J** et **B** et entre -3500 et +3500 pour la matrice **E**. Ces valeurs d'initialisation étant déterminées empiriquement.

5.4. Evolution des filtres

La figure n°2 représente l'évolution du taux d'erreur de regroupement du meilleur individu de la population de filtres sur les bases d'évolution et de cross validation.

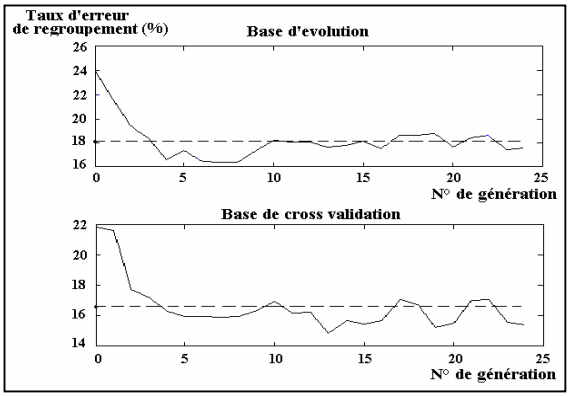


Figure 2 : Evolution des performances des filtres de prétraitement. Trait plein: meilleur individu de la population. Trait discontinu: système de référence (sans prétraitement).

On peut observer que les performances des filtres dépassent rapidement celle du système de référence sur la base de cross validation. Le filtre sélectionné à l'issus de cette phase d'évolution est celui qui présente les meilleures performances sur la base de cross validation (génération n°13).

5.5. Résultats obtenus

L'évaluation des performances du filtre de prétraitement obtenu a été effectuée sur les bases de développement (DEV2) et de test (TEST2) de la campagne d'évaluation ESTER phase 2. La première base est composée de 8 heures d'enregistrements radiophoniques. La seconde base est celle utilisée pour l'évaluation des soumissions de la campagne. Cette base est également composée de 8 heures d'enregistrements audio. La particularité de cette base est la présence de deux heures d'enregistrements provenant de deux radios inconnues. Le tableau n°1

présente les résultats obtenus sur ces bases avec le système de référence et le système intégrant le filtre de prétraitement élaboré.

Tableau 1 : Taux d'erreur de regroupement sur les bases DEV2 et TEST2.

	Base DEV2	**Bases TEST2**
Système de référence	**17.38%**	**21.52%**
Système avec prétraitement	**15.77%**	**21.59%**

Les résultats font apparaître une amélioration significative des performances sur la base DEV2. Cependant, cette amélioration ne se répercute pas sur la base TEST2 où les performances du système proposé sont comparables à celles du système de référence. Ceci peut s'expliquer par le fait que cette base intègre deux heures de signaux provenant de deux nouvelles radios non incluses dans les autres bases.

7. CONCLUSION

Nous avons proposé dans cet article d'utiliser un algorithme génétique pour la synthèse d'un prétraitement non linéaire de la parole adapté au problème de la segmentation et du regroupement du locuteur. Le système ainsi élaboré nous a permis d'obtenir des résultats prometteurs, apportant une amélioration significative des performances sur notre base de développement. Cependant, il apparaît que la méthodologie proposée doit être améliorée pour augmenter la capacité de généralisation du système.

Les systèmes dynamiques non linéaires de part leur complexité et leur richesse permettent d'envisager de nouvelles méthodes de traitement du signal. Nos perspectives de recherche s'orientent vers l'étude de la capacité de ce type de système à l'extraction de caractéristiques discriminantes pour la reconnaissance du locuteur.

8. REMERCIMENTS

Nous tenons à remercier particulièrement Guillaume Gravier pour nous avoir fournis les outils logiciels de segmentation et de regroupement du locuteur présentés dans cet article.

BIBLIOGRAPHIE

[1] B. Gas, JL. Zarader, C. Chavy, M. Chetouani. Discriminant neural predictive coding applied to phonem recognition. *Neurocomputing*, 56:141-166, 2004.

[2] Pitsikalis V. and Maragos, P. Speech analysis and feature extraction using chaotic models. In *Proc. Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 533-536, 2002

[3] Lindgren A.C. Johnson M.T. and Povinelli, R.J. Speech recognition using reconstructed phase space features. In *Proc. Con.f Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 60-63, 2003

[4] Chin-Teng L. Hsi-Wen N. and Jiing-Yuan H. GA-based noisy speech recognition using two-dimensional cepstrum. In *Proc. Conf. Intl. IEEE Transactions on Speech and Audio Processing.*volume 8, pages 664-675, 2000.

[5] Demirekler, M. and Haydar, A. Feature selection using genetics-based algorithm and its application to speaker identification. In *Proc. Con.f Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 329-332, 1999.

[6] C. Charbuillet, B. Gas, M. Chetouani and J.L. Zarader. Filter bank design for speaker diarization based on genetic algorithms. To be pub. In *Proc. Conf. Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[7] Gravier G. and Betser M., Audioseg: Audio Segmentation Toolkit, 2005 *http://www.irisa.fr/metiss/guig/index-en.html*

[8] B. Cessac and J-A Sepulchre, Stable resonances and signal propagation in a chaotic network of coupled units (2004). *Physical Review E,* volume 70, 056111, 2004.

[9] Jaeger, H, Adaptive nonlinear system identification with echo state networks, in *S. T. S. Becker & K. Obermayer, eds, Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, MA*, page. 593-600, 2003.

[10] F. Pasemann, U. Dieckmann, and U. Steinmetz, Evolving Structure and Function of Neurocontrollers, In *Proc. Conf. Congress on Evolutionary Computation Journal*, IEEE Press US, Piscataway, page. 1937-1978, 1999.

[11] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News, *Proceedings of Eurospeech/Interspeech'05*, pages 1149-1152, 2005.

Influence de la corrélation entre le pitch et les paramètres acoustiques en reconnaissance de la parole

G. Cloarec, D. Jouvet & J. Monné

France Telecom – Division R&D – TECH/SSTP
2 avenue Pierre Marzin, 22307 LANNION, France
gwenael.cloarec@francetelecom.com

ABSTRACT

In this paper we compare the role played by the pitch frequency on speaker independent speech recognition performances for two tasks: an isolated word recognition task and a continuous speech recognition task. While introducing pitch and/or voicing directly into the acoustic vector leads to significant improvements on the isolated word recognition task, this method does not bring any improvement on the continuous speech recognition task. On the contrary, modelling the pitch frequency independently of the acoustic parameters leads to small but similar improvements on the two tasks. Those results could be explained by the fact that the improvement brought when introducing pitch and voicing directly into the acoustic vector is related to the correlation between the pitch frequency and the acoustic parameters. This correlation is much less important in the case of the continuous speech recognition task in which prosody can lead to very different pitch values depending on the prosodic context.

1. INTRODUCTION

Bien qu'étant des caractéristiques fondamentales du signal de parole, les paramètres relatifs à l'onde glottique, comme le pitch ou le voisement, ne sont que très rarement utilisés dans les systèmes de reconnaissance de la parole. Ceci est principalement dû au fait qu'ils sont considérés comme étant trop dépendants du locuteur. Cependant l'utilisation du paramètre de pitch est essentielle dans le cas des langues tonales, pour lesquelles le ton est utilisé pour la distinction lexicale. Le pitch peut par ailleurs être introduit pour améliorer la détection de fin de parole dans des conditions difficiles comme dans [1].

On peut néanmoins penser que l'utilisation de paramètres auxiliaires tels que le pitch pourrait améliorer le processus de reconnaissance en permettant, par exemple, de faciliter la distinction entre sons voisés et non voisés. En fait plusieurs études ont déjà été menées afin d'étudier les différents moyens d'intégrer ce coefficient au sein des systèmes de reconnaissance. La méthode la plus simple pour prendre en compte le pitch est de l'intégrer directement au sein du vecteur acoustique. Les résultats obtenus avec cette méthode montrent qu'il existe une différence de comportement entre les tâches de reconnaissance de mots isolés et celles de reconnaissance de parole continue. Ainsi dans [2], le pitch fut introduit au sein du vecteur acoustique et seule une légère amélioration fut obtenue sur de la reconnaissance de parole continue en Mandarin. De même, dans [3], l'utilisation du pitch n'amena qu'une faible amélioration sur de la reconnaissance de parole continue bien qu'une analyse linéaire discriminante ait été appliquée au vecteur acoustique afin de tirer profit de l'information apportée par le pitch, et dans [4] l'utilisation directe d'un indicateur de voisement ne conduisit qu'à de faibles améliorations, toujours sur de la reconnaissance de parole continue. Au contraire, dans [5] nous avons montré que l'utilisation du pitch et/ou du voisement pouvait améliorer significativement la reconnaissance de mots isolés.

Dans cet article nous étudions l'impact du pitch sur deux tâches de reconnaissance de natures différentes; une de reconnaissance de mots isolés, et une autre de reconnaissance de parole continue, et comparons les résultats obtenus.

Le papier est organisé de la façon suivante. La partie 2 décrit les conditions expérimentales. Dans la partie 3 nous reviendrons sur le paramètre de pitch ainsi que sur ses modélisations possibles. La partie 4 présentera les résultats expérimentaux obtenus ainsi que leur analyse. Finalement, les conclusions seront tirées dans la partie 5.

2. CONDITIONS EXPÉRIMENTALES

3.1. Généralités

Le système de reconnaissance utilisé est basé sur des modèles de Markov cachés. La modélisation acoustique prend en compte l'influence contextuelle des phonèmes et repose sur une modélisation gaussienne (8 gaussiennes par densité). L'analyse acoustique est réalisée par l'intermédiaire de l'algorithme Font-End ETSI ES 202 212. Les vecteurs acoustiques de référence sont composés de 33 coefficients : 10 coefficients mel-cepstraux (MFCCs) et le logarithme de l'énergie, ainsi que leurs dérivées temporelles première et seconde.

3.2 Bases de données

Les expériences ont été réalisées sur deux tâches de reconnaissance en mode indépendant du locuteur. La première, que nous appellerons *Communes* par la suite, est une tâche de reconnaissance de mots isolés. Elle est utilisée dans le cadre d'un annuaire téléphonique en environnement RTC, et est basée sur un vocabulaire d'environ 40 000 mots correspondant aux localités françaises [6]. Dans les résultats présentés par la suite, on n'utilise pas de modèle langage, tous les mots du vocabulaire sont donc équiprobables. La seconde, appelée *Plan Resto*, est une tâche de reconnaissance de parole

continue utilisée dans le cadre d'un service de renseignements touristiques avec le système de dialogue décrit dans [7]. Elle est basée sur un vocabulaire de 2200 mots et est utilisée en environnement RTC [8]. Le modèle de langage est basé sur un modèle bi-gramme

Le corpus de test de la tâche *Communes* est composé de 10571 données valides, c'est-à-dire de données correspondant au vocabulaire, et de 1635 données hors vocabulaire, donc à rejeter. Le corpus de test de la tâche *Plan Resto* comprend quant à lui 7507 phrases. Dans les 2 cas, il s'agit de parole spontanée recueillie dans le cadre d'expérimentation de services vocaux.

Les performances sont données en terme de *Word Accuracy*. Pour la tâche *Communes* les performances seront toutes données pour un taux de rejet à tort de 10%, afin de comparer les différentes modélisations du pitch pour un point de fonctionnement donné.

3. MODÉLISATION DU PITCH

3.1 Calcul du pitch

Le pitch est calculé par l'intermédiaire de l'algorithme *extended advanced Front-End (XAFE)* ETSI ES 202 212 [9, 10]. Cet algorithme a été développé afin de fournir une analyse acoustique robuste au bruit dans le cadre de la reconnaissance de parole distribuée. Le paramètre de pitch est utilisé pour pouvoir traiter le cas des langues tonales et pour permettre la reconstruction du signal de parole.

Dans les expériences décrites par la suite, le pitch est représenté par la valeur de la fréquence fondamentale pour les portions voisées du signal de parole. Pour les trames non voisées, sa valeur est arbitrairement fixée à zéro.

Afin de réduire la dépendance au locuteur, nous utilisons également une valeur normalisée du pitch, P_{norm} :

$$P_{norm}(n) = \frac{P(n)}{\frac{1}{N}\sum_{i=1}^{N} P(i)}$$

où le facteur de dépendance au locuteur est approximée par la moyenne du pitch sur les trames voisées. Nous travaillons également avec le logarithme du pitch et de sa valeur normalisée. La valeur de ces paramètres est fixée arbitrairement à -2 pour les trames non voisées.

3.2 Modélisation du Pitch

Intéressons nous maintenant aux différentes façons d'intégrer le pitch au sein de systèmes de reconnaissance de la parole.

Les HMMs utilisent une estimation de la probabilité d'émission du vecteur acoustique, x_t, émis sur l'état s_n:

$$p(x_t | s_n) \quad (1)$$

Si on introduit le pitch, la probabilité d'émission devient :

$$p(x_t, y_t | s_n) \quad (2)$$

Où y_t représente le pitch.

Si on considère le pitch comme un coefficient acoustique supplémentaire, l'équation (2) peut s'exprimer sous la forme suivante:

$$p(x_t, y_t | s_n) = p(x_t \,\&\, y_t | s_n) \quad (3)$$

Où *&* signifie que x_t et y_t appartiennent au même vecteur acoustique.

Par ailleurs, en développant l'équation 2, on obtient :

$$p(x_t, y_t | s_n) = p(x_t | y_t, s_n)\, p(y_t | s_n) \quad (4)$$

En considérant que le pitch est indépendant des paramètres acoustiques et qu'il est modélisé de façon gaussienne, l'équation (4) devient :

$$p(x_t, y_t | s_n) = p(x_t | s_n)\, p(y_t | s_n) \quad (5)$$

Où $p(y_t|s_n)$ est une densité de probabilité mono ou multi gaussienne(s).

4. RÉSULTATS & DISCUSSION

4.1 Pitch dans le vecteur acoustique

Nous considérons ici les résultats obtenus en introduisant directement au sein du vecteur acoustique le paramètre de pitch pour les différentes normalisations présentées dans la partie 3.1. La probabilité d'émission du vecteur acoustique correspond alors à celle définie dans l'équation (3).

La figure 1 présente les résultats obtenus sur la tâche *Communes*. Il apparaît clairement que l'introduction du pitch améliore les performances de reconnaissance puisque dans tous les cas le *Word Accurracy* est plus élevé que celui obtenu avec le modèle de référence. Il passe, par exemple, de 59,40% avec le modèle de référence à 61,17% en utilisant la valeur brute du pitch, soit une réduction relative de 3% du taux d'erreur. Ceci correspond en fait à une réduction relative de 6% du taux de substitution et de 4% du taux de fausse alarme, pour un taux de rejet à tort inchangé. L'utilisation d'une valeur normalisée du pitch ou d'une représentation logarithmique de celui-ci n'apporte pas de réelles améliorations par rapport à l'utilisation de sa valeur brute.

Les résultats obtenus sur la tâche *Plan Resto* sont aussi représentés sur la figure 1. Contrairement au cas précédent, aucune amélioration n'est obtenue par rapport au modèle de référence, quel que soit le paramètre utilisé (valeur brute ou normalisé, échelle linéaire ou logarithmique). Pour illustrer cette observation on peut remarquer que la meilleure performance est obtenue avec la valeur normalisée du pitch, P_{norm}. Le *Word Accurracy* est alors de 72,67 % alors qu'il est de 72.59 % avec le modèle de référence, soit une réduction relative de seulement 0,1%.

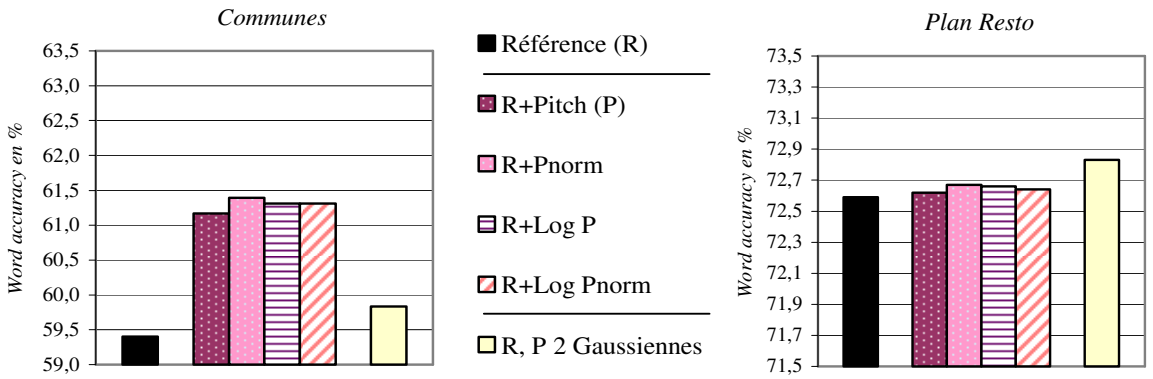


Figure 1 : Performances obtenues sur les tâche *Communes* et *Plan Resto*

4.2 Corrélation entre pitch et MFCCs

Les résultats obtenus précédemment montrent clairement une différence de comportement entre les deux tâches étudiées. En effet, l'introduction directe du pitch au sein du vecteur acoustique permet bien d'améliorer les performances de la tâche de reconnaissance de mots isolés, *Communes*. Au contraire ils n'ont quasiment aucune incidence sur la tâche de reconnaissance de parole continue, *Plan Resto*. Ceci est en adéquation avec les différents résultats présentés dans la littérature.

La table 1 représente la matrice de corrélation du logarithme de l'énergie, des coefficients mel-cepstraux et du pitch pour le phonème /œ/ calculée sur les données d'adaptation de la tâche *Communes*. La corrélation entre le pitch et les coefficients acoustiques est représentée dans la dernière colonne et dans la dernière ligne. Les valeurs élevées (supérieures à 0,20) de corrélation sont présentées en gras. On peut observer que pour un certain nombre de MFCCs (C11 et C9 par exemple), la corrélation avec le pitch est bien plus importante qu'avec les autres paramètres acoustiques.

Table 1 : Matrice de corrélation entre LogE, MFCCs et pitch calculée sur les données de la tâche *Communes* - Phonème /œ/

	LogE	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Pitch
LogE	1.00	-0.21	-0.23	-0.25	-0.01	-0.04	0.09	0.05	0.01	-0.05	-0.06	**0.23**
C2	-0.21	1.00	0.08	-0.00	0.23	-0.07	-0.22	0.10	0.28	-0.05	0.21	**-0.26**
C3	-0.23	0.08	1.00	0.40	-0.23	-0.12	-0.18	-0.11	0.20	0.19	0.18	**-0.34**
C4	-0.25	-0.00	0.40	1.00	-0.25	-0.24	-0.04	-0.10	0.03	0.09	0.28	**-0.24**
C5	-0.01	0.23	-0.23	-0.25	1.00	0.04	-0.05	0.12	0.08	-0.25	0.05	-0.03
C6	-0.04	-0.07	-0.12	-0.24	0.04	1.00	0.17	-0.06	-0.16	0.14	-0.27	**0.31**
C7	0.09	-0.22	-0.18	-0.04	-0.05	0.17	1.00	0.05	-0.36	0.29	-0.16	**0.27**
C8	0.05	0.10	-0.11	-0.10	0.12	-0.06	0.05	1.00	0.12	-0.01	0.35	**-0.22**
C9	0.01	0.28	0.20	0.03	0.08	-0.16	-0.36	0.12	1.00	-0.11	0.33	**-0.56**
C10	-0.05	-0.05	0.19	0.09	-0.25	0.14	0.29	-0.01	-0.11	1.00	-0.07	-0.02
C11	-0.06	0.21	0.18	0.28	0.05	-0.27	-0.16	0.35	0.33	-0.07	1.00	**-0.45**
Pitch	**0.23**	**-0.26**	**-0.34**	**-0.24**	-0.03	**0.31**	**0.27**	**-0.22**	**-0.56**	-0.02	**-0.45**	1.00

La table 2 représente la matrice de corrélation pour le même phonème /œ/ calculée sur les données d'adaptation de la tâche *Plan Resto*. On peut remarquer que le nombre de coefficients de corrélation supérieurs à 0,20 est bien moins important que précédemment : seulement 3 contre 9 pour la tâche *Communes*. Ceci tendrait à montrer que la corrélation entre le pitch et les coefficients cepstraux est moins importante dans le cas de la tâche *Plan Resto*.

Afin de vérifier cette hypothèse, nous avons calculé la corrélation entre le pitch et les différents paramètres acoustiques pour l'ensemble des voyelles communes aux deux tâches. Pour chaque phonème nous avons ensuite relevé le nombre de fois où le coefficient de corrélation entre le pitch et les MFCCs était supérieur à 0,20. Les résultats obtenus sont donnés dans la figure 2 page suivante. On remarque clairement que dans la plupart des cas le nombre de coefficients de corrélation supérieur à 0.20 est plus important pour la tâche *Communes* que pour la tâche *Plan Resto*. Cela signifie que la corrélation entre le pitch et les MFCCs est bien plus grande dans le cas de la tâche de reconnaissance de mots isolés que dans le cas de la tâche de reconnaissance de parole continue. Dans ce dernier cas la prosodie joue un rôle important et peut mener à des variations importantes de pitch selon le contexte prosodique des syllabes. Ceci pourrait expliquer la corrélation moins importante entre le pitch et les MFCCs.

Table 2 : Matrice de corrélation entre LogE, MFCCs et pitch calculée sur les données de la tâche *Plan Resto* - Phonème /œ/

	LogE	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Pitch
LogE	1.00	0.07	-0.04	-0.13	-0.16	-0.20	-0.15	-0.08	0.22	-0.02	0.20	-0.00
C2	0.07	1.00	-0.09	-0.27	0.32	0.04	-0.14	-0.08	0.25	-0.10	0.07	-0.13
C3	-0.04	-0.09	1.00	0.44	-0.19	-0.10	-0.03	-0.28	0.00	0.05	0.08	-0.10
C4	-0.13	-0.27	0.44	1.00	-0.33	-0.20	0.29	-0.16	-0.27	0.08	0.26	-0.07
C5	-0.16	0.32	-0.19	-0.33	1.00	0.12	-0.23	0.04	0.24	-0.32	-0.04	-0.04
C6	-0.20	0.04	-0.10	-0.20	0.12	1.00	-0.09	-0.07	0.04	0.15	-0.29	0.19
C7	-0.15	-0.14	-0.03	0.29	-0.23	-0.09	1.00	0.12	-0.40	0.35	0.09	**0.21**
C8	-0.08	-0.08	-0.28	-0.16	0.04	-0.07	0.12	1.00	-0.05	0.01	0.15	0.11
C9	0.22	0.25	0.00	-0.27	0.24	0.04	-0.40	-0.05	1.00	-0.37	0.03	**-0.39**
C10	-0.02	-0.10	0.05	0.08	-0.32	0.15	0.35	0.01	-0.37	1.00	-0.13	**0.24**
C11	0.20	0.07	0.08	0.26	-0.04	-0.29	0.09	0.15	0.03	-0.13	1.00	-0.19
Pitch	-0.00	-0.13	-0.10	-0.07	-0.04	0.19	**0.21**	0.11	**-0.39**	**0.24**	-0.19	1.00

4.3 Modélisation indépendante du Pitch

Nous avons également modélisé le pitch indépendamment des autres paramètres acoustiques par des distributions à 1, 2, 4 et 8 gaussiennes. La probabilité d'émission correspond cette fois à celle définie dans l'équation (5). Les meilleures performances furent obtenues avec une modélisation à 2 gaussiennes. Les résultats obtenus sur les deux tâches avec cette modélisation sont présentés sur la figure 1. Ici une légère amélioration des performances est observée dans les deux cas. Notons toutefois que pour

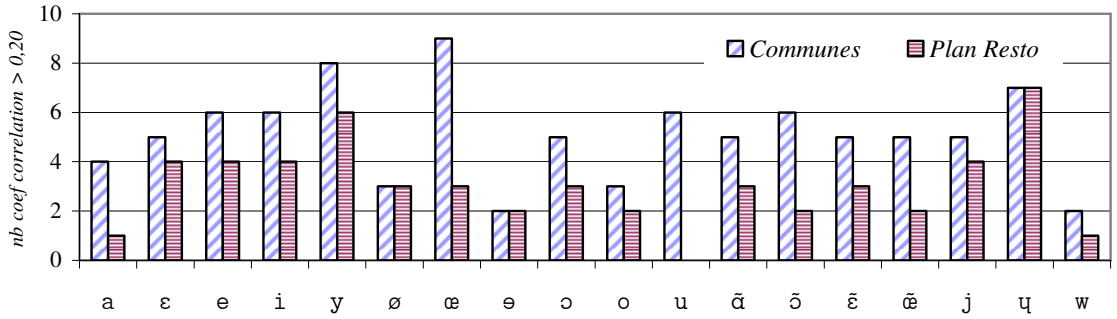


Figure 2 : Comparaison de la corrélation entre le pitch et les coefficients acoustiques pour les tâches *Communes* et *Plan Resto*

la tâche *Communes* cette amélioration est loin d'atteindre celle obtenue précédemment (Le *Word Accuracy* obtenu est de 59,84 % alors qu'il est de 61,17% lorsque le pitch est directement intégré au vecteur acoustique).

Ceci montre néanmoins que les deux tâches de reconnaissance de mots isolés et de parole continue se comportent de façon similaire vis-à-vis du pitch lorsqu'on ne peut tirer profit d'une certaine corrélation entre le pitch et les MFCCs. Ceci montre donc que c'est bien la corrélation plus ou moins importante entre le pitch et les autres paramètres acoustiques qui peut expliquer les différences de résultats entre les deux tâches *Communes* et *Plan Resto* obtenues précédemment en intégrant directement le paramètre de pitch au sein du vecteur acoustique.

5. CONCLUSION

Dans ce papier nous avons comparé le rôle joué par le pitch sur les performances de deux tâches de reconnaissance indépendante du locuteur. En introduisant directement le pitch au sein du vecteur acoustique, les performances de la tâche de reconnaissance de mots isolés ont été nettement améliorées, alors que celles de la tâche de reconnaissance de parole continue sont restées quasiment inchangées. Au contraire en modélisant le pitch indépendamment des autres coefficients acoustiques, nous avons obtenu un comportement similaire des deux tâches de reconnaissance. Ces résultats pourraient s'expliquer par le fait que l'amélioration apportée en intégrant directement le pitch au vecteur acoustique soit liée à la corrélation entre le pitch et les coefficients cepstraux. Cette corrélation est plus importante dans le cas de la tâche *Communes*, qui est une tâche de reconnaissance de mots isolés, que dans le cas de la tâche *Plan Resto*, tâche de reconnaissance de parole continue dans laquelle la prosodie peut mener à des variations importantes du pitch selon le contexte prosodique.

BIBLIOGAPHIE

[1] A. Martin & L. Mauuary, "Voicing Parameter and Energy-Based Speech/Non-Speech Detection for Speech Recognition in Adverse Conditions", in *Proc. Eurospeech'2003, European Conf. on Speech Communication and Technology*, Genève, Suisse, 1-4 Sept. 2003.

[2] S. Liu, S. Doyle, A. Morris & F. Ehsam, "The effect of fundamental frequency on Mandarin speech recognition", in *Proc. ICSLP'98, Int. Conf. on Spoken Language Processing*, Sydney, Australie, vol. 6, pp 2647-2650, 30 Nov.–4 Dec. 1998.

[3] A Ljolje, "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling", in *Proc ICSLP'2002*, *Int. Conf. on Spoken Language Processing*, Denver, USA, pp 2137-2140, 16-20 Sept. 2002.

[4] M. Graciarena, H. Franco, J. Zeng, D. Vergyri, A. Stolke, "Voicing feature integration in SRI's decipher LVCSR System", in *Proc. ICASSP'2004*, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 1, pp 921-924, 17-21 May 2004.

[5] G.Cloarec, J.Monné & D.Jouvet, "Introducing Pitch and Voicing Parameters into Speaker-Independent Speech Recognition Systems", in *Proc. SPECOM'2005*, *10th Int. Conf. on Speech and Computer*, Patras, Grèce, pp95-98, 17-19 Oct. 2005.

[6] Denis Jouvet, K. Bartkova, L. Delphin-Poulat, A. Ferrieux, X. Lamming, J. Monné & C. Raix, "About improving recognition of spontaneously uttered French city names", in *Proc. ICASSP'2003, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong-Kong, vol. 1, pp 544-547, April 2003.

[7] M.D Sadek, A. Ferrieux, A. Ozannet, P. Bretier, F. Panaget, J. Simonin, "Effective Human-Computer cooperative spoken dialogue: the AGS demonstrator", in *Proc. ICSLP'96*, *Int. Conf. on Spoken Language Processing*, Philadelphie, USA, vol. 1, pp 546-549, 3-6 Oct. 1996

[8] C. Raymond, F. Béchet, N. Camelin, R. de Mori, G. Damnati, "Semantic interpretation with error correction", in *Proc. ICASSP'2005, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphie, USA, vol. 1, pp 29-32, March 2005.

[9] ETSI ES 202 212 V1.1.1 (STQ); *Distributed speech recognition; Extended advanced front-end feature extraction*.

[10] A. Sorin, T. Ramabadran, D. Chazan, R. Hoory, M. McLaughlin, D. Pearce, F. CR Wang & Y. Zhang, "The ETSI Extended Distributed Speech Recognition (DSR) Standards: client side processing and tonal language recognition evaluation", in *Proc. ICASSP'2004, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 1, pp 53-56, 17-21 May 2004.

Transformation linéaire discriminante pour l'apprentissage des HMM à analyse factorielle

*Fabrice Lefèvre*et Jean-Luc Gauvain*
Groupe Traitement du Langage Parlé
LIMSI-CNRS, France
{lefevre,gauvain}@limsi.fr

ABSTRACT

Factor analysis has been recently used to model the covariance of the feature vector in speech recognition systems. Maximum likelihood estimation of the parameters of factor analyzed HMMs (FAHMMs) is usually done via the EM algorithm. The initial estimates of the model parameters is then a key issue for their correct training. In this paper we report on experiments showing some evidence that the use of a discriminative criterion to initialize the FAHMM maximum likelihood parameter estimation can be effective.

The proposed approach relies on the estimation of a discriminant linear transformation to provide initial values for the factor loading matrices. Solutions for the appropriate initializations of the other model parameters are presented as well. Speech recognition experiments were carried out on the *Wall Street Journal* LVCSR task with a 65k vocabulary. Contrastive results are reported with various model sizes using discriminant and non discriminant initialization.

1. INTRODUCTION

Ces dernières années, un renouveau d'intérêt a été observé pour la modélisation de la covariance dans les systèmes de reconnaissance de la parole à base de HMM [1, 3, 10, 8]. L'utilisation de matrices de covariance pleines, bien que souhaitable en principe, augmente considérablement en pratique le nombre de paramètres des modèles et complique l'estimation des paramètres. C'est pourquoi des covariances à matrices diagonales sont généralement utilisées dans les systèmes. L'analyse factorielle fournit une stratégie de modélisation intermédiaire qui permet d'obtenir des matrices de covariances pleines avec un moins grand nombre de paramètres. Un modèle générateur de la parole, basé sur un schéma de filtrage statistique du signal, est utilisé. Dans ce modèle, l'hypothèse est faîte que les observations de parole sont le résultat d'une transformation linéaire bruitée à partir d'un espace d'états de plus faible dimension. Les modèles à analyse factorielle ont été récemment généralisés dans le contexte des modèles markoviens [8]. Le HMM à analyse factorielle (ou FAHMM) est un modèle linéaire gaussien basé sur un processus d'évolution d'états constant par morceaux. Les vecteurs d'états sont générés par un HMM utilisant des mixtures de gaussiennes à covariance diagonale. Comme le montre [8], les FAHMM fournissent une cadre général intégrant beaucoup d'autres modèles standards de covariances comme l'analyse factorielle partagée (SFA) [10], l'analyse factorielle indépendante (IFA) [1] ou encore les modèles à covariances semi-liées (STC) [3]. Avec les FAHMM, différents niveaux de partage de paramètres peuvent être envisagés conduisant à différents degrés de complexité des composants statistiques.

L'utilisation optimale des FAHMM repose sur la configuration d'un nombre important de paramètres. Tout comme les HMM traditionnels, la taille des vecteurs d'observation ainsi que le nombre de gaussiennes par état doivent être fixés. De plus, dans le cas des FAHMM, la taille de l'espace d'état doit être fixé ainsi que le nombre de gaussiennes qui lui sont associées. Les autres paramètres du modèle sont appris en utilisant une procédure de type EM. La qualité des valeurs initiales est alors essentielle pour permettre une bonne convergence. L'approche évaluée dans notre travail tente d'améliorer ce point en introduisant un critère discriminant lors de la sélection des dimensions de l'espace d'état. Pour cela nous proposons d'obtenir les dimensions de l'espace d'état à partir d'une transformation linéaire discriminante héteroscedastique (*Heteroscedastic Linear Discriminant Analysis*, HLDA) [5].

Les expériences ont été menées sur une tâche de reconnaissance de la parole continue à grand vocabulaire, *Wall Street Journal* [7]. Une comparaison est faîte entre les HMM à matrices de covariance diagonales et pleines et les FAHMM appris avec les initialisations standard ou discriminante.

2. HMM À ANALYSE FACTORIELLE

Les FAHMM sont une généralisation à espace d'état dynamique des systèmes à analyse factorielle à composants multiples. Le modèle générateur utilisé dans les FAHMM pour chaque indice de temps t et chaque état j est décrit par les équations suivantes :

$$o_t = C_j x_t + v_t \quad (1)$$

$$x_t \sim \sum_k c_{jk}^{(x)} \mathcal{N}(x_t; \mu_{jk}^{(x)}, \Sigma_{jk}^{(x)})) \quad (2)$$

$$v_t \sim \sum_l c_{jl}^{(o)} \mathcal{N}(v_t; \mu_{jl}^{(o)}, \Sigma_{jl}^{(o)})) \quad (3)$$

avec o_t un vecteur d'observation de dimension n, x_t un vecteur d'état de dimension p et v_t un vecteur d'erreur d'observation de dimension n. Toute les matrices de covariance sont diagonales. La structure interne de la covariance est capturée par la matrice C_j, appelée la matrice

*F. Lefèvre est aussi avec l'équipe Dialogue Homme-Machine du LIA-Université d'Avignon

de chargement des facteurs. La distribution d'un vecteur d'observation o_t pour un état donné j, une composant de l'espace d'état k et une composante du bruit d'observation l, est obtenue par intégration sur le vecteur d'état x_t. La distribution qui en résulte est une gaussienne, dont le vecteur moyenne et la matrice de covariance sont donnés par :

$$\mu_{jkl} = C_j \mu_{jk}^{(x)} + \mu_{jl}^{(o)} \quad (4)$$

$$\Sigma_{jkl} = C_j \Sigma_{jk}^{(x)} C_j' + \Sigma_{jk}^{(o)}. \quad (5)$$

La densité de probabilité conditionnelle des observations sur les états des FAHMM peut être vue comme une mixture de densités gaussiennes avec $M^{(o)}M^{(x)}$ matrices de covariance pleines, dont les moyennes et les matrices de covariances sont données par les équations (4) et (5). La vraisemblance marginale d'une observation considérant l'état j est alors obtenue en sommant sur les deux ensembles de gaussiennes. Ce calcul nécessite d'inverser $M^{(o)}M^{(x)}$ matrices de covariance pleines de taille $n \times n$.

Une présentation détaillée des formules des reestimation EM ainsi que les conditions générales de l'apprentissage des FAHMM peut être trouvée dans [8]. Dans notre système, ces formules de réestimation ont été intégrées avec la nuance qu'une segmentation constante est utilisée pour l'apprentissage EM (*apprentissage de Viterbi*). Afin de réduire l'effort de calcul durant l'apprentissage des modèles, l'algorithme à deux niveau, a été adopté : une boucle d'itération interne plus rapide permet d'accélérer la convergence.

L'initialisation des paramètres des modèles est un problème important dans le cadre de l'algorithme EM dans la mesure où elle conditionne la possibilité d'atteindre une bonne solution après convergence. Pour les FAHMM, un point initial raisonnable consiste à convertir un HMM standard (avec des composants mono-gaussiens) en un FAHMM équivalent en utilisant les cepstres statiques comme dimensions de l'espace d'état. Cette proposition fait l'hypothèse que les vecteurs de l'espace d'état sont fortement corrélés avec les dimensions statiques des vecteurs d'observation cepstraux, ce qui n'est pas forcement très réaliste. Pour cette raison, nous proposons d'utiliser une projection discriminante de type HLDA pour l'initialisation des modèles.

3. INITIALISATION DISCRIMINANTE

Des travaux récents sur la modélisation acoustique [5, 3, 9] ont conduit à une adoption généralisée de la technique de transformation linéaire discriminante HLDA dans les systèmes de reconnaissance de l'état-de-l'art. L'objectif de la transformation discriminante des observations consiste à trouver un espace de projection de faible dimension mais conservant l'information discriminante. HLDA est une méthode basée sur le maximum de vraisemblance pour estimer une projection linéaire des vecteurs d'observation de dimension n vers un sous-espace de dimension p. Comme pour LDA, le sous-espace obtenu doit permettre une meilleure séparation des classes, chaque classe étant modélisée par une gaussienne. HLDA généralise LDA en éliminant le recours à une matrice unique comme matrice de covariance intra-classe, ce qui conduit à un meilleur espace de projection lorsque les classes sont héteroscedastiques.

En pratique, la transformation linéaire A est divisée en 2 sous-matrices : A_p transforme l'espace original vers un sous-espace de dimension p et A_{n-p} vers le sous-espace des dimensions rejetées. L'hypothèse faîte est que les moyennes et les variances des n-p dimensions rejetées sont représentées par les dimensions correspondantes des moyennes et variances transformées globales. L'optimisation de la fonction objective associée à HLDA peut être obtenue par des méthodes numériques (comme le gradient conjugué) ou en utilisant une procédure d'optimisation par maximum de vraisemblance réalisée ligne par ligne [3]. Dans cette dernière approche (utilisée dans notre expérience), chaque ligne de la matrice de transformation est mise à jour de façon séquentielle en utilisant le vecteur de cofacteurs de la ligne et la projection des paramètres actuels du modèle.

Lorsque HLDA est appliquée aux HMM pour la reconnaissance de la parole, de bons résultats sont généralement obtenus en utilisant les états liés des HMM comme classes pour HLDA, chacune étant représentée par une gaussienne à matrice de covariance pleine [9]. Plusieurs initialisations sont envisageables pour la matrice de projection. Même si une matrice identité est la solution la plus simple, nous avons observé des résultats légèrement meilleurs en utilisant le ratio de Fisher ou une solution LDA (cette dernière est utilisée par défaut dans nos expériences).

L'usage classique de la technique HLDA consiste, après avoir réduit de manière optimale un espace de grande dimension en un espace plus petit possédant de bonnes propriétés discriminantes, à de construire les nouveaux modèles acoustiques dans ce nouvel espace. Dans notre proposition, la transformation HLDA est utilisée pour définir un espace d'état discriminant pour les FAHMM. Pour ce faire, l'espace d'état est défini par les dimensions utiles de la projection HLDA.

En combinant la définition du sous-espace HLDA $x_t = A_p o_t$ avec le modèle FAHMM donné par l'équation (1), nous observons que la matrice pseudo-inverse de A_p (une matrice rectangulaire $p \times n$) peut être une bonne solution pour initialiser les matrices de chargement de facteurs. Pour nos expériences, la matrice inverse de Moore-Penrose a été utilisée [2], $\hat{C}_j = A_p^+$.

Une fois la matrice de chargement définie, les mixtures de gaussiennes des espaces d'état et de bruit d'observation doivent être initialisées. Elles sont dérivées directement de l'espace HLDA. Les paramètres peuvent alors être obtenus par un apprentissage des modèles dans l'espace HLDA ou directement en transformant les paramètres de l'espace des observations.

Avec la distribution des vecteurs d'états donnée par (2), les paramètres de la distribution par la premier méthode (*apprentissage*) sont obtenus, pour chaque état j, par un apprentissage EM operé sur le sous-espace d'observation résultant de la projection HLDA

$$\mu_{jk}^{(x)} = A_p \mu_{jk} \quad (6)$$

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p W_{jk} A_p^T) \quad (7)$$

avec μ_{jk} et W_{jk} le vecteur moyenne et la matrice de covariance totale des données originales associées à l'état j pour la gaussienne k. Dans la seconde méthode (*projection*), les vecteurs moyens suivent toujours l'équation (6)

mais les covariances deviennent

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p \Sigma_{jk} A_p^T) \quad (8)$$

avec μ_{jk} et Σ_{jk} obtenus par un apprentissage dans l'espace des observations.

A partir des statistiques de l'espace des observations complet, les valeurs initiales pour les paramètres du bruit d'observation sont obtenues en retirant la projection des paramètres de l'espace d'état des paramètres a priori de l'espace des observations. Une covariance diagonale est suffisante a priori dans la mesure où aucune compensation des éléments non-diagonaux n'est réalisée. Lors de l'initialisation, les mixtures d'observation et de bruit doivent avoir le même nombre de gaussiennes ($M^{(o)}$) et la distribution de l'espace d'état est mono-gaussienne. Sinon, l'association entre les composants des observations et de l'espace d'état serait indéfinie et complexe à établir.

Avec la distribution des vecteurs de bruit d'observation définie par l'équation (3), l'initialisation par *apprentissage* conduit à :

$$\mu_{jl}^{(o)} = \mu_{jl} - \hat{C}\mu_{j1}^{(x)} \quad (9)$$

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\Sigma_{j1}^{(x)}\hat{C}^T) \quad (10)$$

et la méthode par *projection* modifie les covariances selon :

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\mathrm{diag}(A_p \Sigma_{j1} A_p^T)\hat{C}^T) \quad (11)$$

Dans ce contexte, un grand soin doit être apporté à appliquer un seuil aux variances du bruit d'observation de sorte à pouvoir calculer les gaussiennes résultantes.

4. DESCRIPTION DU CORPUS ET DU SYSTÈME

Les expériences on été menées dans le cadre de la tâche de dictée vocale de textes, représentée par le corpus *Wall Street Journal* [7], et les conditions de test correspondent à l'évaluation ARPA Hub3 de 1995. Les données acoustiques d'apprentissage ont été prononcées par 355 locuteurs pour un total de 100 heures de parole. Le test Hub3 consiste en parole lue en studio par 20 locuteurs pour un total de 45 minutes. L'analyse acoustique produit des paramètres cepstraux à partir de l'échelle de fréquence Mel estimée sur la bande 0-8kHz toutes les 10ms. La moyenne cesptrale est retirée aux paramètres. Le vecteur acoustique d'observation de 39 dimensions est composé de 12 coefficients cepstraux avec la log-énergie, complétés par leur dérivée des premier et deuxième ordres.

Le système de reconnaissance de dictée vocale du LIMSI utilise des modèles de Markov cachés à densités continues avec des mélanges de gaussiennes pour la modélisation acoustique et des modèles linguistiques de type n-grammes estimés sur de grands corpus de textes. Chaque modèle phonétique dépendant du contexte est un HMM gauche-droit à états liés obtenus grâce à un arbre de décision.

La reconnaissance est opérée en trois étapes : 1) génération d'une hypothèse initiale, 2) adaptation des modèles et génération d'un graphe de mots, 3) adaptation des modèles et génération de l'hypothèse finale. L'hypothèse initiale est utilisée pour l'adaptation des modèles acoustiques à l'aide de la technique du MLLR [6] préalablement à la génération du graphe de mots. Un modèle de langage de type 3-grammes avec back-off est utilisé lors des deux premières étapes. L'hypothèse finale est engendrée à partir d'un modèle 4-grammes et des modèles acoustiques adaptés lors de la seconde étape.

TAB. 1: Taux d'erreur en mots (WER %) et nombre de paramètres par état (η) pour les systèmes HMM à matrices de covariance diagonales et pleines. $M^{(o)}$ est le nombre de gaussiennes par état.

	$M^{(o)}$	*1*	*8*	*16*	*32*
Diagonales	η	78	624	1248	2496
	WER	13.8	8.8	8.5	8.0
	$M^{(o)}$	*1*	*2*	*4*	*8*
Pleines	η	819	1638	3276	6552
	WER	10.9	9.5	9.2	10.3

L'apprentissage EM est réalisé avec une segmentation fixe (*apprentissage de Viterbi*). Les modèles acoustiques dépendants du genre sont obtenus par une adaptation MAP des modèles indépendants du genre [4]. Le jeu de modèles acoustiques comprend 4k phones en contexte, inter-mots et dépendants de la position dans le mot, comptant pour 9k états liés. Le vocabulaire de reconnaissance contient 65k mots avec 77k prononciations phonétiques. Les modèles de langage 3 et 4-grammes résultent de l'interpolation linéaire de modèles appris sur différents ensemble de données (transcriptions des données acoustiques d'apprentissage, journaux...). Un graphe de prononciation est associé à chaque mot permettant des prononciations alternatives.

5. RÉSULTATS EXPÉRIMENTAUX

Afin de définir une référence, une configuration de modèles HMM à matrices de covariance diagonales est évaluée. Tous les résultats rapportés ont été obtenus avec des modèles dépendant du genre du locuteur et adaptés de façon non supervisée au locuteur. Le tableau 1 donne les taux d'erreur en mots et le nombre de paramètres libres par états (η) pour quatre valeurs de $M^{(o)}$ (le nombre de composantes gaussiennes par état). On constate que le taux d'erreur décroît de manière constante avec le nombre de paramètres, jusqu'à atteindre 8% avec 32 gaussiennes par état.

Une expérience contrastive a été menée avec des HMM à base de covariances pleines. Les résultats obtenus avec de 1 à 8 gaussiennes par état sont donnés dans la seconde partie du tableau 1. Le taux d'erreur le plus bas est 9.2% avec 4 gaussiennes par état, c'est à dire environ 1% au dessus du meilleur résultat obtenu avec des matrices de covariance diagonales. Ce résultat est cohérent avec d'autres résultats rapportés pour de la parole conversationnelle [11], et il justifie le développement d'une modélisation intermédiaire.

Le tableau 2 présente les taux d'erreur pour plusieurs configurations de FAHMM avec le nombre de paramètres par état. La taille de l'espace d'état a été fixée à 13 avec un espace d'observation de 39 dimensions. Dans le cas de l'initialisation standard des paramètres, un ensemble de HMM à mixtures de gaussiennes avec matrices de covariance diagonales est utilisé pour initialiser les FAHMM comme proposé dans [8]. Une seule matrice de charge-

TAB. 2: Taux d'erreur en mots ($WER\%$) et nombre de paramètres par état (η) pour les 3 configurations des FAHMM (toutes avec $n = 39$ et $p = 13$) : standard, initialisations HLDA apprise et par projection. $M^{(x)}$ et $M^{(o)}$ sont respectivement le nombre de composants dans les espaces d'etat et d'observation.

$M^{(x)}$	$M^{(o)}$	*1*	*2*	*3*	*4*	*5*	*6*
		Initialisation standard					
6	η	715	793	871	949	1027	1105
	WER	9.2	8.9	8.8	8.9	9.1	8.8
		Init. HLDA (projection)					
	WER	9.0	8.5	8.5	8.4	8.6	8.3
		Init. HLDA (apprentissage)					
	WER	9.2	8.9	8.6	8.8	8.5	8.3
3	η	637	715	793	871	949	1027
	WER	9.5	9.4	9.0	8.7	8.7	8.3
1	η	585	663	741	819	897	975
	WER	11.0	10.0	9.9	9.2	9.4	9.2

ment de facteurs est utilisée par état, elle est partagée entre toutes les gaussiennes. Les deux premières lignes du tableau correspondent à l'initialisation standard avec 6 gaussiennes pour l'espace d'état et de 1 à 6 gaussiennes pour l'espace d'observation. Le meilleur taux d'erreur 8.8% est obtenu avec $M^{(o)} = 6$. Bien que le nombre de parametres indépendants par état reste peu élevé (1105) en comparaison des modèles à covariances diagonales, le coût de décodage est significativement plus important. Ceci explique qu'il est difficile d'aller au delà de la configuration 6×6.

Les résultats de l'initialisation discriminante des FAHMM sont donnés dans les deux lignes suivantes du tableau 2, pour l'initialisation HLDA par apprentissage d'une part et par projection d'autre part. Lorsque que le nombre de composants dédiés au bruit est augmenté, les performances s'améliorent plus rapidement pour la méthode *par projection*, bien que au final les 2 approches obtiennent des résultats comparables pour $M^{(o)} = 6$ (8.3%). Avec l'initialisation discriminante, le taux d'erreur est diminué de 0.5% pour la meilleure configuration ($M^{(x)}$=6, $M^{(o)}$=6) comparée à l'initialisation standard. Avec environ un millier de paramètres indépendants, le taux d'erreur est aussi inférieur à celui obtenu avec des modèles à matrices de covariance diagonales.

La dernière partie du tableau donne des résultats supplémentaires pour l'initialisation de type *apprentissage* en utilisant moins de composants pour l'espace d'état (3 et 1). Avec $M^{(x)} = 3$, les résultats sont meilleurs que ceux observés pour $M^{(x)} = 6$ à nombre de paramètres fixe. Finalement, les performances baissent avec $M^{(x)} = 1$. Ces résultats tendent à montrer que l'équilibre entre les valeurs de $M^{(x)}$ and $M^{(o)}$ est un point délicat pour obtenir de bonnes performances avec les FAHMM.

CONCLUSION

Les HMM à analyse factorielle ont été appliqués sur une tâche de reconnaissance de parole continue grand vocabulaire, utilisant 100 heures de données d'apprentissage. Le système intègre un modèle de langage 4-grammes appris sur un vocabulaire de 65k mots et des adaptations supervisée (MAP) et non-supervisée (MLLR) des modèles acoustiques FAHMM. Une méthode a été proposée pour améliorer l'initialisation des dimensions de l'espace d'état pour l'apprentissage des FAHMM, basée sur un critère discriminant (HLDA).

A la suite d'une série d'expériences, nous avons observé que les FAHMM appris selon l'approche proposée présentent des performances légèrement inférieures à celles des HMM à matrices de covariance diagonales (8.3% vs 8.0%) mais supérieures à celles des HMM à covariances pleines (8.3% vs 9.2%) et des FAHMM standards (8.3% vs 8.8%). De plus, si nous fixons les nombre de paramètres indépendants du système vers 1k par état, les FAHMM avec une initialisation discriminante donnent alors de meilleurs résultats que tous les autres modèles, y compris les modèles à covariance diagonale.

RÉFÉRENCES

[1] H. Attias. Independent factor analysis. *Neural Computation*, 11(4) :803–851, 1999.

[2] S. Campbell and C. Meyer. *Generalized Inverses of Linear Transformations*. Dover Publications, New-York, 1991.

[3] M. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3) :272–281, 1999.

[4] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, 1994.

[5] N Kumar and A Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4) :283–297, 1998.

[6] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9 :171–185, 1995.

[7] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the ICSLP*, pages 899–902, Banff, 1992.

[8] A-V.I. Rosti and M.J.F. Gales. Factor analysed hidden markov models for speech recognition. *Computer Speech and Language*, 18(2) :181–200, 2004.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proceedings of the IEEE ICASSP*, Istanbul, 2000.

[10] L. Saul and M. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2) :115–125, 2000.

[11] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The ibm 2004 conversational telephony system for rich transcription. In *Proceedings of the IEEE ICASSP*, volume I, pages 205–208, Philadelphia, 2005.

Proposition d'une nouvelle méthodologie pour la sélection automatique du vocabulaire d'un système de reconnaissance automatique de la parole

Brigitte Bigi

CLIPS/IMAG Lab. CNRS,
BP 53, 38041 Grenoble cedex 9, France
Tél. : ++33 (0)4 76 51 45 26 - Fax : ++33 (0)4 76 63 55 52
Mél : {brigitte.bigi}@imag.fr

ABSTRACT

The vocabulary of an Automatic Speech Recognition (ASR) system is a significant factor in determining its performance. The goal of vocabulary selection is to construct a vocabulary with exactly those words that are the most likely to appear in the test data. This paper proposes a new measure to evaluate the quality of a vocabulary regarding a domain-specific ASR application. This Q_α-measure is based on the trade off between the target lexical coverage and vocabulary size. Experiments were carried out on French Broadcast News Transcriptions using the Q_α-measure compared to the state-of-the-art method. Results of these two methods favor systematically the proposed methodology.

INTRODUCTION

Dans le cadre d'une amélioration des performances des systèmes de Reconnaissance Automatique de la Parole (RAP), nos travaux visent à développer une méthodologie pour sélectionner efficacement le vocabulaire du système. La sélection du vocabulaire est un processus appliqué à un ou plusieurs corpus de différentes sources, origines ou périodes, dont le résultat est la définition d'une liste de mots. Le but est de déterminer la combinaison optimale des vocabulaires produits par les différents corpus.

La couverture lexicale indique le taux des mots du vocabulaire présents dans le corpus de test. Le problème de la construction d'un vocabulaire consiste à obtenir la meilleure couverture lexicale, sur un corpus de développement. Les méthodes manuelles de construction du vocabulaire consistent à sélectionner les K mots les plus fréquents observés sur des corpus d'apprentissage. Les valeurs de K sont définies manuellement en fonction de la taille, la période ou la nature des corpus disponibles.

Une méthode automatique qui permet d'ordonner les mots à partir de plusieurs corpus a été proposée dans [4] puis dans [1]. Le principe consiste en une interpolation linéaire des modèles unigrammes estimés sur des corpus d'apprentissage, de manière à minimiser la perplexité du modèle interpolé sur le corpus de développement. Le vocabulaire choisi contient les K mots les plus probables, K étant fixé manuellement. La méthode présentée dans cet article repose sur la proposition d'une nouvelle mesure de qualité qui intègre la couverture lexicale et une notion de pertinence relative à la taille du vocabulaire. Cette mesure s'intègre dans une démarche de recherche du meilleur rapport entre le taux de mots hors-vocabulaire et la taille du vocabulaire.

En section 1, nous montrons une étude qui met en avant la problématique relative à la sélection du vocabulaire et argumentons sur la nécessité de développer une méthode automatique de sélection. La méthodologie proposée est décrite en section 2. Elle repose sur la proposition d'une nouvelle mesure de qualité d'un vocabulaire. Le principe général de cette approche consiste à utiliser les corpus disponibles pour créer un ensemble de vocabulaires possibles et d'utiliser la mesure Q_α pour les comparer. La section 3 expérimente la méthodologie de sélection sur les données du projet ESTER dont le corpus se compose du journal "Le Monde", de transcriptions d'émissions radiophoniques, auxquelles nous avons ajouté des pages téléchargées sur des sites de radios et journaux sur l'Internet. Dans la dernière section, les résultats obtenus sont confrontés à la méthode automatique couramment utilisée dans le domaine [4, 1]. Le gain obtenu par la méthode proposée concerne la réduction systématique du taux de mots hors-vocabulaire sur les corpus de développement et de test.

1. PROBLÉMATIQUE

1.1. État de l'art

Le problème de la sélection du vocabulaire est notamment abordé dans [3]. L'auteur montre que la taille du vocabulaire d'un système de RAP a deux effets. D'une part, le taux de mots hors vocabulaire (MHV) est réduit, aidant à faire moins d'erreurs de substitutions dues aux mots inconnus. D'un autre côté, les entrées lexicales ajoutées augmentent la confusion acoustique sur les mots, induisant de nouvelles erreurs de reconnaissance. Il conclut que la taille optimale du vocabulaire dépend de la tâche pour laquelle le système est dédié et du système lui-même. Cependant, il ne propose pas de solution pour obtenir ce vocabulaire optimal.

Dans [4], trois méthodes sont proposées pour obtenir une liste de mots triés par priorité décroissante, à partir de plusieurs corpus. La meilleure des trois méthodes repose sur la maximisation de la vraisemblance de l'estimation des comptes, à partir de m corpus. Le but est de trouver les λ_m coefficients de l'interpolation linéaire des m modèles unigrammes. De la même manière, [1] propose de calculer un jeu de coefficients d'interpolation des unigrammes des différents corpus. Ils vérifient l'hypothèse que minimiser la perplexité du corpus de développement calculée avec un modèle interpolé implique la minimisation du taux de MHV du vocabulaire construit à partir de ce modèle. Dans cette méthode, il reste à l'utilisateur à choisir la valeur de K, taille du vocabulaire final.

1.2. Description des corpus

Les données sur lesquelles nous travaillons dans cet article proviennent de la campagne ESTER qui vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques. La campagne est organisée dans le cadre du projet EVALDA sous l'égide scientifique de l'Association Francophone de la Communication Parlée avec le concours du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et de ELDA (Evaluations and Language resources Distribution Agency).

TAB. 1: Description des corpus d'apprentissage

Source	Période	$\|W\|$	$\|V\|$	MHV
Le Monde	1987-2003	413M	855K	0,21
Transcriptions	1998-2000,2003	1M	34K	2,26
Web	2003-2004	2,5M	52K	3,16

Deux corpus écrits ont été utilisés pour la campagne ESTER Phase 2. Le premier est un corpus audio manuellement transcrit. Ces transcriptions proviennent principalement de France-Inter, France-Info, Radio France International et Radio Télévision Marocaine. Le second est un corpus de textes du journal "Le Monde". Le corpus des transcriptions a été divisé en 3 parties pour l'apprentissage, le développement et les tests. Le corpus de développement concerne l'année 2003 ; il contient 97K occurrences pour un vocabulaire de 9800 mots. Le corpus de test concerne les mois d'octobre et décembre 2004 ; il contient 119K occurrences pour un vocabulaire de 11317 mots. Enfin, nous avons ajouté un corpus web qui provient de l'aspiration quotidienne de données ciblées du web durant la période de juin 2003 jusqu'à avril 2004 (de 20 à 200 pages web par jour, de radios et journaux). Le tableau 1 indique la période concernée par chacun des corpus d'apprentissage, le nombre total de mots $|W|$, la taille du vocabulaire complet $|V|$, et le pourcentage de mots hors-vocabulaire sur le corpus de développement.

1.3. Couverture lexicale : discussion

Sélectionner un vocabulaire est une tâche difficile dans la mesure où elle présuppose des besoins de l'application. Dans la plupart des cas, le vocabulaire "complet" est impossible à atteindre, ou alors, il implique que celui-ci soit d'une très grande taille, bien supérieure à la limite imposée par le système auquel il est dédié. Notons $V = \{w_1, w_2, \cdots, w_K\}$ l'ensemble des K mots d'un vocabulaire V ; notons également $c(w, C)$ le nombre d'occurrences du mot w dans le corpus C. La couverture lexicale L d'un vocabulaire V sur un corpus C s'écrit :

$$L(V, C) = \frac{\sum_{w \in V} c(w, C)}{\sum_{w \in C} c(w, C)}$$

Plus le vocabulaire est grand, plus la couverture lexicale est élevée. On utilise de façon identique le pourcentage de MHV, calculé tel que : $MHV = 100 - (L(V, C) \times 100)$.

La figure 1 indique le taux de MHV du corpus du journal "Le Monde", divisé par années, sur un corpus de transcriptions de 2003. Elle montre la pertinence d'avoir des données récentes.

La figure 2 indique le taux de MHV du corpus du journal "Le Monde", divisé par années, sur un corpus de transcriptions de 2003, en regroupant les années. Elle montre le rôle négligeable de la quantité de données : une seule année suffit pour obtenir une bonne couverture lexicale, et l'ajout de données plus anciennes n'apporte rien, voire dégrade la qualité du vocabulaire.

La figure 3 indique la couverture lexicale des corpus de transcription, "Le Monde" et du Web. Elle met en exergue l'importance d'avoir des données de la même origine que celles de l'application visée.

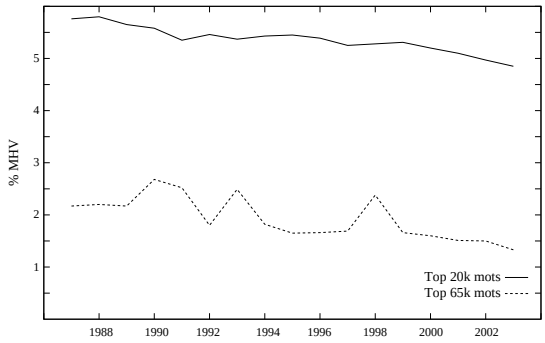


FIG. 1: MHV des vocabulaires du journal "Le Monde", selon l'année du corpus

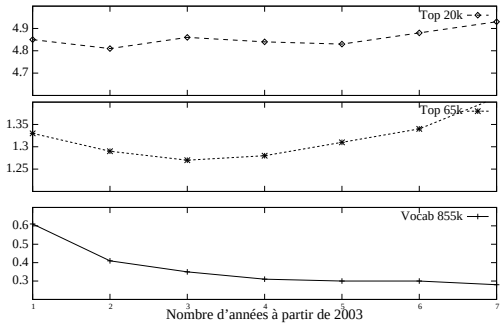


FIG. 2: MHV des vocabulaires du journal "Le Monde", en augmentant les données chaque année

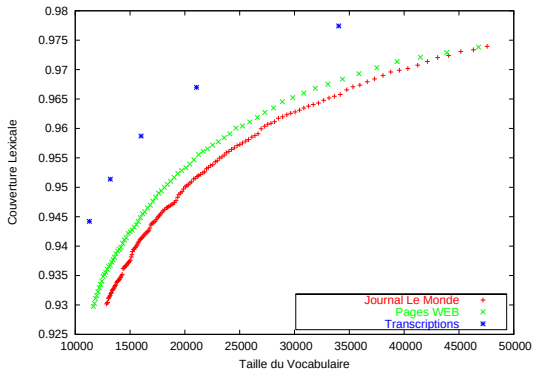


FIG. 3: Relation entre couverture lexicale, taille du vocabulaire et origine du corpus

2. MÉTHODE DE SÉLECTION AUTOMATIQUE

2.1. Production de vocabulaires candidats

La sélection d'un vocabulaire s'appuie sur l'utilisation de la fréquence ; les mots sont triés par ordre décroissant de leur nombre d'apparitions et le choix se porte sur les K mots les plus fréquents. Cette méthode, bien que très largement utilisée, pose un problème de coupe "brute" car des mots peuvent avoir la même fréquence autour de la valeur de K. C'est classiquement le rang des mots dans l'ordre alphabétique qui détermine les mots choisis. Nous proposons que les vocabulaires candidats se limitent à ceux qui ont des mots de fréquence supérieure à N. Par exemple, le vocabulaire du corpus des transcriptions contient 34037 mots dont 21059 ont plus d'une occurrence ; ainsi on choisira soit un vocabulaire de 34037 mots, soit de 21059 mots, mais on ne choisira pas une taille intermédiaire car les 12978 mots de différence ont la même fréquence (égale à 1). Pour un corpus donné, on dispose de plusieurs vocabulaires candidats, en faisant varier N.

On peut également segmenter les grands corpus selon des périodes de temps, ou de thèmes, selon l'application à laquelle le système est dédié. De nombreuses possibilités existent lorsque l'on dispose de corpus de tailles, origines et périodes différentes. Dans ce cas, il est souhaitable de segmenter au maximum les corpus.

Tous ces vocabulaires candidats, obtenus directement à partir des corpus, ou sous-corpus, peuvent être combinés (union ou intersection), afin d'obtenir de nouveaux vocabulaires candidats. Il reste alors à définir une mesure de qualité pour les comparer et décider du vocabulaire final de l'application.

2.2. Sélection par une mesure de qualité

La définition de la mesure proposée nécessite l'introduction des notations suivantes :
- M_c est la masse des mots communs, i.e. la somme des occurrences dans le corpus $\mathcal{C}$, de tous les mots présents à la fois dans le vocabulaire V et dans le corpus $\mathcal{C}$;
- M_f est la masse des mots exclus, i.e. la somme des occurrences dans le corpus $\mathcal{C}$, de tous les mots hors du vocabulaire V qui sont présents dans le corpus $\mathcal{C}$;
- M_i est la masse des mots ignorés, i.e. le nombre de mots présents dans le vocabulaire mais absents du corpus.

Selon cette notation $M_c + M_f$ représente la somme de toutes les occurrences de tous les mots de $\mathcal{C}$. La couverture lexicale L d'un vocabulaire V sur un corpus $\mathcal{C}$ se réécrit comme suit :

$$L(V,\mathcal{C}) = \frac{M_c}{M_c + M_f}$$

Dans cet article, nous introduisons une "mesure de pertinence du vocabulaire" qui tient compte des mots du vocabulaire qui ne sont pas utiles. Cette pertinence, notée R pour *relevance* en anglais, s'évalue telle que :

$$\mathcal{R}(V,\mathcal{C}) = \frac{M_c}{M_c + M_i}$$

La mesure Q_α combine la couverture lexicale L et la pertinence R, suivant le même principe que la F_α-mesure, largement utilisée en recherche d'information, par exemple, pour combiner le rappel et la précision des systèmes. Q_α permet la sélection des V_α vocabulaires dont la taille augmente au fur et à mesure qu'α croît. Q_α se calcule comme suit :

$$Q_\alpha = \frac{e^{\frac{\alpha+1}{2}} \times \mathcal{R}(V,\mathcal{C}) \times L(V,\mathcal{C})}{\left(e^{\frac{\alpha+1}{2}} \times \mathcal{R}(V,\mathcal{C})\right) + L(V,\mathcal{C})}$$

3. EXPÉRIMENTATION

Cette section détaille l'utilisation de la méthodologie proposée, sur les données décrites en section 1.2.

3.1. Vocabulaires candidats

La première étape de la méthodologie consiste à définir un ensemble de vocabulaires candidats. Pour cette expérimentation, nous avons fait les choix suivants :
- "Le Monde", avec N variant de 50 à 1800 par pas de 5 (soit 351 vocabulaires possibles) ;
- Transcriptions, avec N variant de 0 à 8 par pas de 1 (soit 9 vocabulaires possibles) ;
- Web, avec N variant de 0 à 50 par pas de 1 (soit 51 vocabulaires possibles).

ainsi que toutes les unions de 2 ou 3 de ces vocabulaires.

3.2. Sélection par la mesure Q_α

La table 2 montre les vocabulaires choisis par la mesure Q_α parmi les milliers de vocabulaires candidats. Nous avons fait varier α de 1 à 9, ne reste à l'utilisateur qu'à choisir celui dont la taille est la plus adaptée à l'application visée (ou au système qui l'utilise). La première colonne indique la mesure utilisée, les colonnes suivantes indiquent le détail des vocabulaires utilisés pour créer le vocabulaire choisi, dont la taille est indiquée dans la dernière colonne.

3.3. Validation

La validation consiste à comparer les taux de MHV des vocabulaires obtenus par notre méthode automatique avec celle proposée dans [4, 1]. Dans [1], il est prouvé que la méthode manuelle donne de moins bons résultats que la méthode automatique. Nous comparerons donc nos résultats directement à celle-ci. Chacun des corpus $\mathcal{C}_i$ est utilisé pour apprendre une distribution de probabilités $P(w,\mathcal{C}_i)$. Le problème consiste à trouver les λ coefficients de l'interpolation linéaire entre ces unigrammes :

$$P(w,\mathcal{C}_1,\cdots,\mathcal{C}_n) = \sum_{i=1}^{n} \lambda_i P(w,\mathcal{C}_i)$$

où $\sum_{i=1}^{n} \lambda_i = 1$. Les meilleures valeurs de λ optimisent la perplexité estimée sur le corpus de développement des transcriptions. Les mots obtenant les K meilleures probabilités $P(w,\mathcal{C}_1,\cdots,\mathcal{C}_n)$ sont sélectionnés pour le vocabulaire final. Pour notre expérimentation, l'interpolation optimale sur le corpus de développement est la suivante :

$$\begin{aligned} P(w,ester) = \; & 0,758 \times P(w,transcriptions) + \\ & 0,139 \times P(w,lemonde) + \\ & 0,103 \times P(w,web) \end{aligned}$$

Les résultats sont présentés dans la table 3, pour le corpus de développement, et dans la table 4 pour le corpus de test. Dans ces deux tables, il est intéressant de noter que pour les 9 vocabulaires proposés par la méthode Q_α, le

TAB. 2: Description des vocabulaires proposées par la méthodologie

Mesure	Transcriptions		Le Monde		Web		Union
	N	$\|V\|$	N	$\|V\|$	N	$\|V\|$	$\|V\|$
Q_1	6	8927	1735	13416	49	4247	15055
Q_2	3	13196	1725	13478	49	4247	17200
Q_3	2	16006	1250	16668			20944
Q_4	1	21059	735	23300			28898
Q_5	1	21059	415	32662			36544
Q_6	1	21059	330	37217			40519
Q_7	0	34037	230	45483			54220
Q_8	0	34037	140	61210			67963
Q_9	0	34037	75	84140			89139

taux de MHV est inférieur à celui de la méthode utilisant l'interpolation linéaire. Même si le gain est parfois négligeable, il n'en est pas moins significatif car systématique (c'est-à-dire pour toutes les valeurs de α).

TAB. 3: Comparaison des taux de MHV (développement)

	% MHV	
$K = \|V\|$	Q_{alpha}	$P(w, ester)$
15055	4,07	4,04
17200	3,51	3,60
20944	2,89	2,95
28898	2,07	2,16
36544	1,59	1,90
40519	1,41	1,59
54220	1,06	1,13
67963	0,88	0,90
89139	0,68	0,71

TAB. 4: Comparaison des taux de MHV (test)

	% MHV	
$K = \|V\|$	Q_{alpha}	$P(w, ester)$
15055	4,82	4,98
17200	4,37	4,45
20944	3,72	3,75
28898	2,70	2,83
36544	2,13	2,51
40519	1,92	2,16
54220	1,47	1,45
67963	1,18	1,21
89139	0,92	1,00

Pour des raisons de clarté de la présentation de cette section, nous avons choisi de ne présenter qu'une seule expérience effectuée sur les trois corpus dont nous disposons. De nombreuses autres expériences ont été menées, notamment en divisant le corpus "Le Monde" par années, ou en regroupant les années les plus récentes. Dans tous les cas, les résultats sont en faveur de la méthode Q_α. A titre d'exemple, le vocabulaire issu de notre participation à la phase 2 du projet ESTER était composé de 21010 mots, créé comme suit :

$$(\text{"}LeMonde\text{"}, 2003, N = 30$$
$$\wedge\ \text{"}LeMonde\text{"}, 2001 - 2003, N = 241)$$
$$\vee\ Transcriptions, N = 2$$

Le taux de MHV sur le corpus de développement est de 2,84 %, contre 2,94 % pour la méthode à base d'interpolation linéaire. De même le taux de MHV sur le corpus de test est de 3,66 %, contre 3,74 % pour l'autre méthode.

TAB. 5: *Description du corpus anglais*

	$\|W\|$	$\|V\|$	MHV
Meetings - Test	16,5K	1943	-
Meetings - Train	1,0M	14669	3,2
Fisher - Train	5,3M	33110	3,3
Broadcast News - Train	131M	229535	2,2

Nous avons également conduit la même expérience sur des données en langue anglaise pour une application de transcriptions de réunions (table 5). Dans cette expérience, on observe un gain pour chacune des valeurs α, de 1 à 9. La mesure Q_6 propose un vocabulaire de 10124 mots avec 3,40% de MHV, tandis que l'interpolation linéaire obtient 3,50% de MHV pour un vocabulaire de même taille.

CONCLUSION

Cet article a proposé une méthode qui peut déterminer entièrement automatiquement le vocabulaire d'un système de RAP. Contrairement à la méthode à base d'interpolation linéaire, la méthode proposée ne nécessite pas de fixer *a priori* la taille du vocabulaire final. De plus, les résultats montrent une réduction systématique des taux de MHV des vocabulaires choisis par Q_α.

RÉFÉRENCES

[1] A. Allauzen and J-L. Gauvain. Construction automatique du vocabulaire d'un système de transcription. In *XXV Journees d'Etudes sur la Parole*, Fès (Maroc), 2004.

[2] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of Interspeech 2005*, 2005.

[3] R. Rosenfeld. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proceedings of Eurospeech 1995*, 1995.

[4] A. Venkataraman and W. Wang. Techniques for effective vocabulary selection. In *Proceedings of Eurospeech 2003*, 2003.

Théorie de la syllabe et durées vocaliques : Vers une interprétation unifiée du rôle de la structure syllabique et de la nature des segments.

Crouzet, O. & Angoujard, J.-P.

Université de Nantes, Nantes Atlantique Universités,
Laboratoire de Linguistique de Nantes – LLING, EA3827,
UFR Lettres et Langages, Chemin de la Censive du Tertre, BP81227, Nantes, F-44000 France.
{olivier.crouzet|jean-pierre.angoujard}@univ-nantes.fr
http://www.lettres.univ-nantes.fr/lling/

ABSTRACT

It is generally agreed that vowel duration may be influenced by both phonetic context and syllable structure. Though it may seem reasonable to call for two different variables in this respect, we show that the *rythm and substance* [1] approach to syllable structure may offer a common framework for unifying these two sources of variation into a single theoretical account. The results of a speech production experiment involving french speakers are described which confirm that this syllabic approach accounts for some of the predicted deviations in vowel duration when syllabic and contextual effects are involved. Though further studies should be conducted, this framework seems particularly promising for the understanding of the relationship between articulatory, phonological and rythmic influences on speech production mechanisms.

1. INTRODUCTION

L'étude des facteurs influençant les durées vocaliques fait ressortir un ensemble de variables prédictives complémentaires extrêmement nombreuses. Plusieurs de ces variables sont liées à une représentation prosodique de la chaîne parlée (positionnement de la voyelle dans les groupes intonatifs, structure de la syllabe *portant* la voyelle). L'effet de la nature des segments environnants constitue également un facteur prédictif important des variations observées. Dans la présente communication, nous focalisons notre approche sur une proposition d'unification des modélisations syllabique et segmentale.

1.1. Facteurs influençant les durées vocaliques

L'influence syllabique La structure de la syllabe constitue une source importante de variation de la durée vocalique [10]. Ainsi, la voyelle initiale [a] est plus longue lorsqu'elle est prononcée dans [aʃ] que dans [aʃa]. Dans le premier cas, la fricative est en position terminale (i.e. *coda*, cf. Fig. 1) de la syllabe alors que dans le second cas, elle est en position initiale (i.e. *attaque*) de la syllabe suivante. Maddieson [10] montre que ces variations de la durée vocalique constituent des indices phonétiques (c'est à dire *observables*) de la structure syllabique des séquences. L'influence de la structure syllabique est actuellement prise en considération dans certains systèmes de synthèse vocale [cf. notamment 11].

L'influence segmentale À cette influence syllabique sur les durées vocaliques, s'ajoute une source contextuelle non-prosodique, c'est à dire indépendante de l'organisation *hiérarchique* des segments dans la chaîne parlée : l'effet du contexte phonétique sur la durée de la voyelle précédente. Dans une séquence VC (Voyelle-Consonne), la voyelle est plus longue si elle est suivie d'une fricative que d'une occlusive (p.ex. [as] vs. [at]) [6, 7], elle est également plus longue si elle est suivie d'une consonne voisée que d'une consonne non-voisée (p.ex.[ad] vs. [at]) [3]. Ces phénomènes pourraient cependant être modulés par, voire confondus avec, d'autres variables comme la vitesse d'élocution. Dans les mécanismes de perception de la parole, ces variations contextuelles peuvent être utilisées comme des indices acoustiques du voisement des consonnes post-vocaliques. Lindblom [9] rend compte de ces variations à partir d'un modèle de la coordination des lèvres et de la mâchoire.

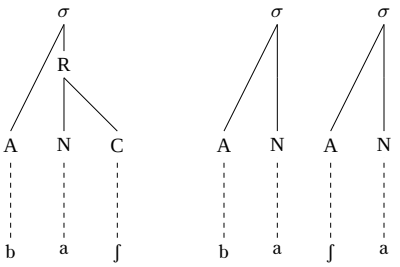


FIG. 1: Représentation syllabique de [baʃ] et [baʃa] selon un cadre théorique classique communément utilisé dans les travaux actuels sur la syllabe. σ représente le *sommet* de la syllabe. R désigne la *rime*, elle-même constituée du *noyau* (N) et de la *coda* (C). L'*attaque* (A) est directement rattachée au sommet de la syllabe.

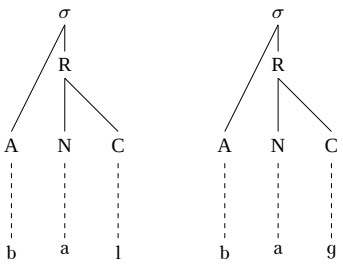


FIG. 2: Représentations syllabiques de [bal] et [bag] dans le cadre théorique classique. La représentation des séquences [bal] et [bag] est strictement identique dans ce cadre-là.

2. LA SYLLABE : RYTHME ET SUBSTANCE

Dans le cadre de la théorie syllabique développée en [1], la syllabe est considérée comme l'instanciation d'une chaîne de segments et d'une grille rythmique (cf. Fig. 3).

Tout segment est caractérisé par un ensemble de propriétés qui déterminent sa nature, c'est à dire sa *substance* [4, 8] et se voit attribuer une position rythmique déterminée à partir de ses caractéristiques substantielles. Toute voyelle est un pic rythmique, toute consonne est un creux rythmique. Ces creux rythmiques peuvent avoir des statuts différents. En autorisant le rattachement d'un segment donné à une position de creux post-syllabique, on lui attribue le statut de *coda* (ici appelée « Position 3 »). Seule une classe particulièrement limitée de segments peut se rattacher à cette position 3 (notamment le /l/). La plupart des segments consonantiques se rattachent nécessairement à une « Position 1 » (c'est à dire à une attaque de syllabe). Le statut des fricatives (et par conséquent du /ʁ/ français) est à ce titre encore sujet à débat dans le cadre de cette approche.

L'interaction entre nature des segments et grille rythmique détermine les caractéristiques de la courbe supérieure, laquelle est en général interprétée comme une représentation de la sonorité [5] mais pourrait également être vue comme une représentation symbolique des alternances d'ouverture et de fermeture du conduit vocal.

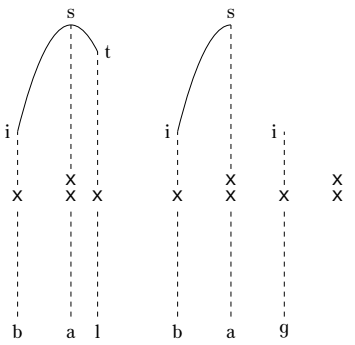


FIG. 3: Représentation syllabique de [bal] et [bag] selon [1]. La ligne inférieure correspond à la séquence des segments ; sur la ligne médiane, on trouve la grille rythmique qui représente l'alternance des position rythmiques (faibles / fortes) de la séquence ; la ligne supérieure (la courbe) est classiquement décrite comme une représentation de l'échelle de sonorité des segments. Ces trois niveaux de représentation fournissent une description particulièrement riche de ce qu'est une syllabe dans ce cadre théorique. Contrairement aux conceptions classiques, [bal] et [bag] ont une représentation syllabique toute à fait différente : [g] ne peut pas être rattaché en position finale de syllabe et se trouve nécessairement à l'attaque d'une syllabe à noyau vide.

Dans ce cadre théorique, les séquences [bal] et [bag] se voient attribuer une représentation profondément différente. Cette différence est du même ordre que celle qui caractérise l'opposition entre les formes [aʃ] et [aʃa] : dans le premier cas, la consonne est en position finale de syllabe, dans le second cas elle se trouve à l'attaque de la syllabe suivante. Ce parallélisme permet de faire des prédictions concernant les variations de durée vocalique en fonction de la structure syllabique *et* de la nature des segments.

Si les représentations syllabiques présentées ici permettent de prédire la réalisation de voyelles plus courtes dans [bal] que dans [bag], c'est que l'on peut voir dans ces représentations syllabiques une conception proche des propositions d'Öhman [12] concernant la coordination des gestes vocalique et consonantique lors de la production de la parole [cf. également 2]. En effet, selon [12] le geste vocalique est un geste *global* ; tout geste consonantique n'est quant à lui qu'un geste *local* et transitoire entre deux gestes vocaliques. Si l'on intègre ces propositions dans la théorie syllabique décrite ici, on peut considérer qu'un geste vocalique s'étale sur toute la largeur de la syllabe alors que les gestes consonantiques auront une portée beaucoup plus réduite. Nous proposons que ces gestes vocaliques sont limités à l'empan de la syllabe (en tant que représentation symbolique de la chaîne parlée). Passer d'une syllabe à une autre consiste à passer d'une voyelle à une autre en traversant des états consonantiques, ces états consonantiques étant rattachés à la position initiale ou finale d'une syllabe en fonction de leur nature. Ces propositions sont représentées dans les Figures 4 à 6.

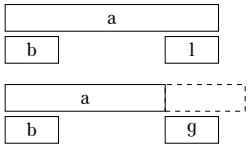


FIG. 4: Représentation temporelle de [bal] et [bag]. Le modèle syllabique développé en [1] permet de proposer une représentation temporelle proche des propositions de [12]. Cette représentation conduit à prédire des différences de durée vocalique en fonction de la nature des segments. Le modèle prédit des variations de la durée des voyelles en fonction de la structure syllabique sur la base des même principes (cf. Fig. 5 et 6).

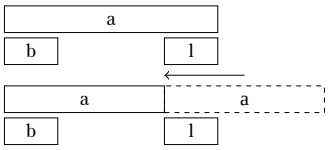


FIG. 5: Représentation temporelle de [bal] et [bala]. La diminution de la durée de la voyelle dans [bala] s'explique par la réduction de l'empan attribué à la voyelle dans cette forme.

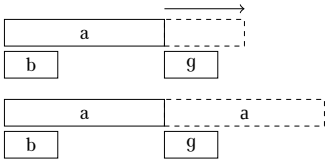


FIG. 6: Représentation temporelle de [bag] et [baga]. L'augmentation de la durée de la voyelle dans [bag] peut s'expliquer ici par l'allongement de l'empan global attribué au geste vocalique (incluant donc aussi la portion correspondant au noyau vide).

3. EXPÉRIENCE

Nous faisons l'hypothèse que la conception rythmique de la syllabe est un modèle particulièrement approprié pour rendre compte des phénomènes phonologiques associés à la syllabe mais aussi des phénomènes articulatoires et acoustiques qui pourraient interagir avec cette représentation.

Afin de valider cette hypothèse, nous avons étudié les variations de la durée des voyelles en fonction de leur contexte consonantique *et* de la structure syllabique prédite par les deux approches théoriques présentées. En effet, si le modèle syllabique classique conduit à considérer les effets syllabiques et contextuels comme deux effets indépendants, l'interaction forte entre représentation syllabique et segmentale mais aussi entre représentation phonologique et contraintes motrices est centrale dans le

modèle rythmique. À partir de cette opposition constitutive des deux modèles, il est possible de faire l'hypothèse suivante. Si l'on compare par exemple les sons [l] et [g], le modèle rythmique prédit une représentation syllabique fondamentalement différente sur la base de leur nature segmentale, ce que ne fait pas le modèle syllabique classique. Dans ce cadre-là, le modèle rythmique prédit des différences de durée vocalique entre les voyelles produites dans [bal] et [bag], ce qui peut évidemment s'expliquer, en dehors du modèle syllabique, par des contraintes articulatoires telles que celles décrites par Lindblom [9].

Mais le modèle rythmique permet également de faire une prédiction tout à fait centrale : si la consonne finale du mot est effectivement rattachée à la syllabe suivante dans des séquences comme [labalɛtɔ̃bɐ] ou [labagɛtɔ̃bɐ], les représentations des consonnes qui nous intéressent sont quant à elles particulièrement différentes lorsqu'elles sont suivies d'une autre consonne ne donnant pas lieu à resyllabation. Ainsi, les mots [bal] et [bag] dans [labaldəmɔ̃fʁɛʁ] et [labagdəmɔ̃fʁɛʁ] ne peuvent-ils être analysés comme structurellement identiques. Cette différence a un impact majeur sur les prédictions que l'on peut faire concernant les durées vocaliques. En effet, s'il est probable que l'on observera des différences de durée entre les séquences de type CV#CV (où # représente une *frontière* syllabique dans le cadre du modèle syllabique classique) et celles de type CVC#C, le modèle rythmique –contrairement au modèle classique– prédit que cet effet devrait être plus fort pour la latérale alvéolaire [l] –qui changera effectivement de position syllabique– que pour les occlusives voisées – qui elles ne peuvent jamais être rattachées à une position de coda, Cf. Fig. 7 et 8 ainsi que Fig. 3.

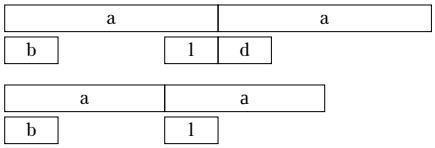


FIG. 7: Représentation temporelle de [balda] et [bala]. La resyllabation du [l] conduit à prédire une réduction particulièrement importante de la durée de la voyelle dans [bala].

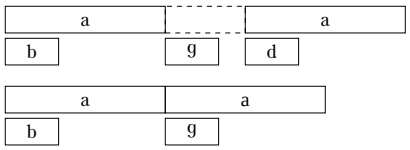


FIG. 8: Représentation temporelle de [bagda] et [baga]. L'absence de chevauchement entre les gestes vocalique et consonantique final dans [bag] conduite à prédire un fonctionnement totalement différent du passage de [bagda] à [baga]. En effet, dans les deux cas, le geste vocalique ne se prolonge pas dans la consonne suivante en raison de la présence d'un noyau vide (cf. Fig. 3). Les variations de durée devraient donc être moins importantes.

3.1. Méthode

Participants Nous présentons ici les résultats de deux locuteurs ayant participé à cette expérience. Nous procédons actuellement à l'enregistrement d'un nombre plus conséquent de locuteurs mais présentons néanmoins les résultats des analyses statistiques inférentielles conduites sur ces données afin de mettre en évidence la stabilité des résultats observés. Les locuteurs sont tous de langue maternelle française et ne présentent aucun trouble avéré de la production de la parole.

Enregistrement Les enregistrements ont été réalisés sur la piste gauche d'un enregistreur DAT Tascam DA-P1 avec un microphone Sennheiser e835. Les locuteurs étaient assis face au microphone dans une pièce calme et lisaient la liste des phrases à un rythme d'une phrase toutes les 2 secondes. Les enregistrements ont ensuite été transférés sur ordinateur par l'intermédiaire d'une carte son Sound-Blaster Audigy LS et digitalisés en monophonique sur 16 bits à 16kHz.

Matériel Le noyau syllabique des mots étudiés correspondait à l'une des 4 voyelles {i,a,y,u} (3 mots différents par voyelle $*$ 6 classes de consonnes post-vocaliques : [ʁ] vs. [l] vs. {v,z,ʒ} vs. {f,s,ʃ} vs. {b,d,g} vs. {p,t,k}). Il s'est avéré impossible de maintenir la consonne initiale constante pour chaque *rime* VC ; les mots ont donc été choisis afin de s'assurer d'un contrebalancement des caractéristiques de la consonne initiale, ceci afin de limiter les effets éventuels de cette consonne sur l'analyse des durées vocaliques : pour chaque ensemble de mots se terminant par une classe de consonnes donnée, on s'est assuré d'une répartition équitable d'occlusives, de fricatives et de nasales, d'une part ; de voisées et de non-voisées d'autre part. Au total, chaque locuteur a produit les 72 mots dans 2 contextes différents, soit 144 réalisations par locuteur.

Mesures acoustiques La mesure des durées vocaliques a été réalisée à l'aide du logiciel Wavesurfer [14] à partir de la courbe de modulation d'amplitude et du spectrogramme à bande large en réglant le contraste pour maximiser l'émergence des formants vocaliques. Afin de réduire la variabilité des mesures effectuées, nous avons choisi *a priori* un ensemble de critères de localisation du début et de la fin d'une voyelle. Ces critères sont les suivants :

– Lorsque la consonne précédente (resp. suivante) est non-voisée, le début (resp. la fin) de la voyelle est localisé au moment où apparaissent (resp. disparaissent) s*imultanément* les formants et l'énergie correspondant au voisement.
– Lorsque la consonne précédente (resp. suivante) est voisée, le début (resp. la fin) de la voyelle est localisé au moment où apparaît la frontière entre bruit de frottement (lié à la friction ou au relâchement) et tracés formantiques (resp. au moment où commence la phase d'occlusion ou de friction).
– En cas d'incertitude, nous avons fait reposer notre décision sur l'écoute du signal et c'est toujours la solution qui tend à réduire la durée de la voyelle qui a été privilégiée.

Si ces critères peuvent de toute évidence être sujet à débat (comme probablement tout critère dans l'analyse des données scientifiques), leur respect nous permet d'atteindre un maximum de régularité dans les décisions, ce qui nous paraît être le plus important pour l'étude de cette hypothèse d'interaction en fonction du contexte et de la structure.

3.2. Résultats

Le Tableau 1 présente les durées moyennes de la voyelle en fonction de la nature de la consonne post-vocalique et du contexte dans lequel elle est produite.

TAB. 1: Durées vocaliques mesurées (en ms.) en fonction de la nature de la consonne finale et de son contexte de production (contexte vocalique donnant lieu à resyllabation vs. contexte consonantique ne donnant pas lieu à resyllabation).

Consonne post-vocalique	Structure CV#CV	Structure CVC#C
ʁ	93.7	94.5
l	97.2	89.7
v,z,ʒ	108.2	101.0
f,s,ʃ	78.4	84.8
b,d,g	75.2	73.3
p,t,k	73.6	71.5

Nous avons conduit une analyse de variance de ces données en évaluant les effets de la nature de la voyelle, de la consonne post-vocalique et de la structure syllabique provoquée par le contexte post-consonantique sur la durée vocalique. On n'observe aucun interaction du timbre de

la voyelle avec les autres facteurs ($F < 1$). L'interaction de 2nd ordre est marginale ($F_{[30,60]} = 1.56, p = 0.07$). Cette variable a donc été retirée des analyses ultérieures. Les effets globaux du type de consonne post-vocalique et de la structure syllabique déclenchée par le contexte de production du mot sont significatifs (respectivement $F_{[5,8]} = 64.20, p < 0.01$ pour l'effet du contexte et $F_{[2,8]} = 14.24, p < 0.01$ pour la structure syllabique). L'interaction entre ces deux variables est également significative ($F_{[10,20]} = 22.29, p < 0.01$).

Nous avons vu rapidement, dans l'introduction (Sec. 2), que les fricatives et le [ʁ] soulèvent de nombreuses questions quant à leur représentation dans le modèle rythmique. Nous nous limiterons donc ici à une analyse comparative de la latérale alvéolaire ([l]) et des occlusives voisées ({b,d,g}) afin d'évaluer la validité des hypothèses formulées (nous laissons de côté les occlusives non-voisées pour ne pas introduire dans la comparaison de modification liée au voisement). À cet effet, nous avons conduit une analyse des contrastes pertinents pour répondre à la question suivante : la différence de durée vocalique observée en fonction de la structure syllabique (CV#CV vs. CVC#C) est-elle plus importante pour la consonne ([l] que pour les occlusives voisées (les données spécifiquement associées à cette question sont illustrées Fig. 9) ?

En structure CV#CV, les voyelles sont plus longues lorsqu'elles sont suivies d'une latérale alvéolaire que d'une occlusive voisée ($p < .05$). En structure CVC#C, on observe une différence de durée vocalique plus importante, laquelle est également significative ($p < .001$).

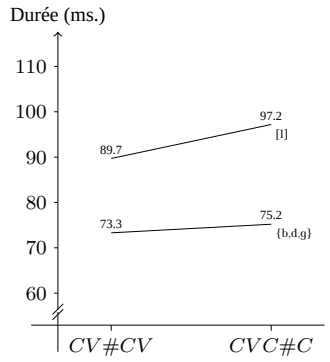


FIG. 9: Comparaison des durées vocaliques mesurées (en ms.) en fonction de la consonne post-vocalique ([l] vs. {b,d,g}) et de son contexte de production (contexte vocalique donnant lieu à resyllabation vs. contexte consonantique ne donnant pas lieu à resyllabation).

4. DISCUSSION

S'il n'est pas possible d'affirmer que nous sommes ici en présence d'une *interaction statistique* entre contexte phonétique de la voyelle et structure syllabique, le fait d'aboutir à un seuil de probabilité nettement plus bas pour les structures CVC#C que pour les structures CV#CV nous semble particulièrement prometteur. En effet, à la lecture de ces analyses, il semble que l'allongement de la durée de la voyelle en fonction de la structure syllabique puisse effectivement être plus important devant [l] que devant une occlusive voisée puisque le test statistique fait *plus facilement* émerger cet effet.

Il conviendra de poursuivre ce travail en comparant différents modes de calcul des durées vocaliques ainsi qu'en ayant systématiquement recours à une transformation des données de durée segmentale en distribution log-normale [13] afin d'accroître la validité statistique des analyses mais aussi de recourir à des statistiques Bayésiennes afin d'évaluer à l'aide d'outils statistiques plus appropriés les modifications d'*ampleur* des effets observés.

Quoi qu'il en soit, les résultats présentés ici constituent une première tentative d'unification de deux sources de variation des durées vocaliques considérées classiquement comme deux phénomènes distincts. Des travaux complémentaires doivent évidemment être conduits afin d'approfondir notre compréhension de ces phénomènes. Cette approche théorique nous semble cependant particulièrement intéressante, et les résultats préliminaires présentés dans cette étude contribuent à renforcer cette position, car elle nous semble constituer un outil conceptuel tout à fait approprié à une réflexion sur les relations entre contraintes articulatoires, représentations phonologiques et mécanismes cognitifs de type rythmique.

RÉFÉRENCES

[1] J.-P. Angoujard. *Théorie de la syllabe*. Paris : CNRS Éditions, 1997.

[2] C. P. Browman and L. Goldstein. Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman, editors, *Papers in Laboratory Phonology I : Between the grammar and physics of speech*, pages 341–376. Cambridge University Press, Cambridge, UK, 1990.

[3] M. Chen. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22 :129–159, 1970.

[4] G. N. Clements. The geometry of phonological features. *Phonology Yearbook*, 2 :225–252, 1985.

[5] G. N. Clements. The role of the sonority cycle in core syllabification. In J. Kingston and M. E. Beckman, editors, *Papers in laboratory phonology I : Between the grammar and physics of speech*, pages 283–333. Cambridge University Press, Cambridge, UK, 1990.

[6] A. House. On vowel duration in English. *The Journal of the Acoustical Society of America*, 33 :1174–1178, 1961.

[7] A. House and G. Fairbanks. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, 25 :105–113, 1953.

[8] J. Kaye, J. Lowenstamm, and J.-R. Vergnaud. The internal structure of phonological elements : A theory of charm and government. *Phonology*, 7(2) :193–231, 1990.

[9] B. Lindblom. Vowel duration and a model of lip mandible coordination. In *Speech Transmission Laboratory – Quarterly Progress and Status Report*, volume 8, pages 1–29. Royal Institute of Technology, Stockholm, 1967. URL `http://www.speech.kth.se/qpsr/pdf/1967/1967_8_4_001-029.pdf`.

[10] I. Maddieson. Phonetic cues to syllabification. In V. Fromkin, editor, *Phonetic Linguistics : Essays in Honor of Peter Ladefoged*, pages 203–221. Academic Press, CA, Orlando, 1985.

[11] R. Ogden, J. Local, and P. Carter. Temporal interpretation in prosynth, a prosodic speech synthesis system. In *Proceedings of the XIV^th International Congress of Phonetic Sciences, ICPhS'99*, 1999.

[12] S. E. G. Öhman. Coarticulation in VCV utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39 :151–168, 1966.

[13] K. M. Rosen. Analysis of speech segment duration with the lognormal distribution : A basis for unification and comparison. *Journal of Phonetics*, 33 :411–426, 2005.

[14] K. Sjölander and J. Beskow. Wavesurfer : An open-source speech tool. In *International Conference on Spoken Language Processing*, Beijing, China, 2000. URL `http://www.speech.kth.se/wavesurfer/`.

Effets aérodynamiques du mouvement du velum : le cas des voyelles nasales du français

Angélique Amelot et Alexis Michaud

Laboratoire de Phonétique et de Phonologie, UMR 7018 CNRS/Sorbonne Nouvelle
19 rue des Bernardins, 75005 Paris
angelique.amelot@univ-paris3.fr, alexis.michaud@univ-paris3.fr

ABSTRACT

Hitherto unpublished data on oral airflow, nasal airflow and velum movement during logatoms read by two French speakers allow for the investigation of the relationships between these three phenomena. There is no straightforward relationship between velar movements and nasal airflow, the latter depending on the relative impedance of both tracts, reflected in the ratio of nasal airflow to oral airflow. The structure of the 168 logatoms is $C_1V_1C_1V_tC_1V_1$, where C_1 = /t/, /d/, /l/, /n/, /s/ or /z/, V_1 = /a/, /i/, /u/ or /y/, and V_t = /ɛ̃/, /ɑ̃/, /ɔ̃/, /a/, i/, /u/ or /y/, allowing for a characterisation of the effect of these consonantal and vocalic contexts on airflow and velum movement. In particular, a hypothesis is put forward concerning the frequent dip below zero of nasal airflow after stop consonants, and its effect on oral airflow.

1. INTRODUCTION

Dans le domaine de la phonétique expérimentale, il existe généralement plusieurs instruments pour aborder un même phénomène ; chacune des méthodes exploratoires possède ses avantages et ses limites, de sorte que les méthodes se complètent mutuellement. Dans le domaine de l'étude des phénomènes linguistiques de nasalité, les limites de l'observation spectrographique ont très tôt conduit à recourir à des mesures aérodynamiques, ainsi qu'à des techniques d'observation du port vélo-pharyngé. Plusieurs études combinent ces deux méthodes exploratoires (notamment Delvaux [7], Amelot [1]), mais sans nécessairement les mettre en regard de façon systématique. Débit d'air nasal et degré d'abaissement du velum ne sont pas déductibles l'un de l'autre, dans la mesure où le débit nasal dépend de l'impédance relative du conduit oral et du conduit nasal (Baken [3, page 408], Ohala [12]). Au-delà de ce constat, il paraît intéressant de mettre en regard le mouvement vélaire avec le *rapport entre débit d'air nasal et débit d'air oral*. Pour des segments qui comportent à la fois un flux d'air nasal et un flux d'air oral, tels que les voyelles nasales du français, il paraît raisonnable de faire l'hypothèse selon laquelle ces deux ensembles de données seraient fortement corrélés. La présente étude doit permettre d'affiner le constat selon lequel le mouvement vélaire serait anticipatoire, tandis que le débit d'air nasal serait tardif et se prolongerait au-delà du segment nasal (Delvaux [6], Basset *et al.* [4]).

2. METHODE

2.1. Corpus et locuteurs

L'étude porte sur 168 logatomes de structure $C_1V_1C_1V_tC_1V_1$, où C_1 = /t/, /d/, /l/, /n/, /s/ ou /z/, V_1 = /a/, /i/, /u/ ou /y/, et V_t = /ɛ̃/, /ɑ̃/, /ɔ̃/, /a/, i/, /u/ ou /y/ (exemple : /tatɛ̃ta/). Les logatomes, placés dans une phrase cadre (« Dites __ trois fois »), ont été répétés deux fois (pour plus de détails, voir Amelot [1]) ; en l'absence de problème technique, c'est la première répétition qui est étudiée ici. Les locuteurs étaient une femme de 29 ans et un homme de 25 ans, locuteurs natifs du français et résidant actuellement en région parisienne.

2.2. Acquisition des données

Les données aérodynamiques ont été enregistrées à l'Hôpital Tenon (par Bernard Roubeau) à l'aide de la station EVA2 (Teston *et al.* [9]). Les données fibroscopiques ont été prises à l'Hôpital Européen Georges Pompidou (par le Dr. Lise Crevier-Buchman) au moyen de la station ATMOS (www.atmosmed.de), qui recueille une image fibroscopique toutes les 40 ms.

2.3. Alignement et mesures

Segmentation et alignement : Des frontières entre segments ont été placées d'après l'examen du signal acoustique et de spectrogrammes ; c'est à cette segmentation que renvoient les expressions telles que « le début de V_t » employées ci-après. Les données enregistrées séparément ont dans un premier temps été alignées par le début acoustique de V_t (voir figure 1) ; dans un second temps, elles ont été alignées par la fin de cette voyelle. Cette procédure repose sur l'hypothèse d'une régularité du comportement du locuteur (en termes de débit de parole, et d'articulation des logatomes) d'un protocole expérimental à l'autre ; cette hypothèse se trouve vérifiée dans une étude comparant des données ainsi alignées à des données prises simultanément (Amelot *et al.* [2]).

Estimation de la hauteur vélaire, Ht : La hauteur du velum a été calculée sur chaque image, par comparaison avec une position d'abaissement maximum (en phase de respiration). Il s'agit d'une mesure relative, en pixels.

Calcul d'un rapport entre débit d'air oral et débit d'air nasal : le débit d'air nasal a été rapporté au débit total d'air égressif (débit oral+débit nasal), suivant la méthode proposée par Krakow [11, page 34].

Calcul d'une hauteur vélaire moyenne au cours de V_t et de la consonne qui suit : la moyenne de Ht a été calculée sur la durée de V_t, à partir des données alignées par le début de V_t, et sur la durée de la consonne suivante à partir des données alignées par la fin de V_t. Il va de soi que ce paramètre à lui seul ne résume pas la courbe de hauteur vélaire, courbe qui elle-même fournit une vue simplifiée du mouvement vélaire. Il a néanmoins été tenté de normaliser ces valeurs, en les transformant en pourcentage de la valeur observée pendant la tenue des occlusives orales (donc pendant une fermeture totale du port vélo-pharyngé).

3. RÉSULTATS

3.1. Analyse qualitative

Les résultats sont très différents selon le contexte consonantique. Les résultats pour les occlusives, plus complexes, seront présentés en détail, suivis de remarques sur les fricatives, la nasale et la liquide.

Les tracés en début de $V_{t\,(nasale)}$ en contexte occlusif (/t/, /d/) paraissent s'organiser selon un continuum entre deux pôles, selon le degré de réalisation de l'occlusion de la consonne qui précède $V_{t\,(nasale)}$ (occlusion à l'intérieur du conduit oral, et fermeture du port vélo-pharyngé).

Pôle 1 : le débit d'air nasal commence après le début acoustique de la voyelle. Un pic de débit oral ressort nettement ; il correspond au relâchement de la consonne (du fait du temps que prend le déplacement de l'air, ce pic se situe dans le premier cinquième de la voyelle). Sa position temporelle coïncide avec un bref passage sous zéro de la courbe de débit nasal, que nous interprétons (à la suite de Benguerel [5, page 109]) comme un effet de l'abaissement du velum juste avant que la cavité nasale n'entre en communication avec la cavité orale. (Voir figure 2.) Dans les cas où ce passage sous zéro coïncide avec le pic de débit d'air oral (par exemple sur la figure 1), il paraît raisonnable de penser que le même mécanisme est à l'œuvre, et qu'il explique l'amplitude plus forte du pic de débit d'air oral.

Pôle 2 : le débit d'air nasal commence avant le relâchement de la consonne ; notre interprétation est que l'occlusion nasale est incomplète. Le débit d'air oral, soit ne comporte pas de pic en début de voyelle (indice de l'absence de l'*augmentation de pression orale, derrière une occlusion orale*, attendue pour une occlusive canonique), soit comporte un pic peu marqué, signe d'une faible explosion. (Ce cas de figure ne peut être présenté visuellement ici faute de place.)

Au plan aérodynamique, la transition entre $V_{t\,(nasale)}$ et la consonne occlusive suivante se caractérise par une chute du débit oral (chute qui se prolonge durant la consonne suivante), et d'un prolongement du débit nasal (avec souvent un second pic, légèrement plus élevé que le premier). Ceci rejoint l'observation de Durand [8, page 210] selon laquelle la nasalité d'une voyelle tend à se propager vers l'occlusive qui suit, et s'arrête avant le relâchement de la consonne.

En contexte fricatif, $V_{t\,(nasale)}$ présente des courbes de débit d'air globalement plus lisses qu'entre occlusives, et doublement symétriques : la courbe de débit nasal et celle de débit nasal présentent une allure inverse l'une de l'autre ; et les courbes en début de syllabe (augmentation du débit nasal, baisse du débit oral) sont relativement symétriques de celles en fin de syllabe (baisse du débit nasal, augmentation du débit oral).

En contexte nasal ou liquide, la variabilité est considérable. Le seul fait qui ressorte avec régularité est que le débit nasal est plus élevé pendant les consonnes nasales que pendant les voyelles nasales, tandis que le débit oral épouse le schéma inverse.

3.2. Analyse quantitative

Les données des deux locuteurs sont présentées séparément, du fait que les valeurs absolues de débit de la locutrice sont très différentes de celles du locuteur ; en théorie, les mesures adimensionnelles (relatives) proposées ici sont directement comparables entre locuteurs, mais il paraît prudent de réserver cette analyse pour un stade auquel un plus grand nombre de locuteurs sera pris en compte.

La proportion de débit d'air nasal pendant V_t est indiquée sur les graphiques 1 (valeurs triées par voyelle) et 2 (par consonne ; les graphiques 2, 3, 5 et 6 présentent seulement les données des logatomes contenant des voyelles nasales). Le graphique 3 représente ce même paramètre pendant la consonne qui suit V_t. Les graphiques 4 et 5 montrent les valeurs de Ht au cours de V_t ; valeurs triées par voyelle sur le graphique 4, par consonne sur le graphique 5. (Rappel : des valeurs élevées de Ht indiquent une position haute du velum.). Le graphique 6 présente les valeurs de Ht pendant la consonne qui suit V_t.

La voyelle /ɔ̃/ a le débit nasal le plus élevé, par rapport au débit oral ; à l'opposé, /ɛ̃/ présente les valeurs les plus basses de ce rapport, alors que son degré d'abaissement du velum est le plus important des sept voyelles. Cette différence aérodynamique paraît due à la différence d'impédance du conduit oral entre ces voyelles. D'autres observations seront évoquées au fil de la discussion.

4. DISCUSSION

La prise en compte du débit oral dans l'analyse des voyelles nasales permet de parvenir à une compréhension affinée de la propagation du débit nasal. La position du velum ne détermine pas à elle seule la répartition de l'air égressif en un flux oral et un flux nasal : cette répartition dépend de l'impédance relative des deux conduits, oral et nasal.

La réalisation canonique des occlusives requiert que le velum soit entièrement remonté, afin que s'établisse une pression intra-orale élevée (Ohala [11]) ; nos données montrent qu'après une voyelle nasale, le velum n'est pas encore tout à fait remonté au moment de l'occlusion orale : tandis que le débit oral devient faible ou nul, une quantité accrue d'air passe par le nez, d'où une augmentation considérable du rapport *débit nasal/débit total*. Le fait est particulièrement net pour les occlusives voisées (voir figure 2) ; le mouvement vélaire aurait ainsi partie liée avec le voisement de la consonne. Hayes [10] indique qu'une légère ouverture du port vélo-pharyngé peut favoriser le voisement d'une consonne, ce qui expliquerait que les consonnes voisées soient plus perméables à la nasalité que les non voisées. Devant V_t, les courbes de débit montrent qu'en cas de forte anticipation de nasalité, une occlusive voisée peut perdre sa phase d'occlusion, et son relâchement (ce que confirment les spectrogrammes).

Les fricatives sont plus perméables à la nasalité que les occlusives, cela bien que la position du velum soit comparable pour ces deux ensembles de consonnes ; le fait s'explique vraisemblablement par le débit d'air élevé au cours des fricatives.

5. CONCLUSIONS

Les principales conclusions sont les suivantes, par ordre croissant d'originalité :

1) Lors de la réalisation d'une occlusive, lorsqu'un pic de débit d'air oral est présent, le débit nasal apparaît après celui-ci ; autrement dit, le pic de débit oral est le signe d'un relâchement de l'occlusion orale précédant l'ouverture du port vélo-pharyngé. A l'inverse, lorsqu'un débit nasal positif est présent dès le début de V_t, il n'est pas observé de pic de débit d'air oral ; ce cas correspond à une réalisation co-articulée, affaiblie, de l'occlusive précédant V_t.

2) Lors de la transition entre une voyelle nasale et une occlusive, le débit d'air oral baisse (signe d'occlusion orale) avant que le port vélo-pharyngé ne soit fermé, d'où un surcroît de débit d'air nasal pendant la consonne (avant qu'intervienne le relâchement de celle-ci).

3) Les voyelles /ɛ̃/ et /ɑ̃/ sont produites avec moins de débit d'air nasal que de débit d'air oral ; le phénomène inverse est observé pour /ɔ̃/, bien que cette dernière possède une *hauteur vélaire intrinsèque* plus élevée.

4) L'apparition du débit nasal s'accompagne d'une diminution du débit oral ; cela est particulièrement net pendant /n/, ce qui suggère que la pression sous-glottique est relativement constante au cours du logatome.

5) En contexte consonantique fricatif (/s/, /z/), débit oral et débit nasal sont inversement corrélés ; en contexte /l/, une telle relation ne ressort pas nettement, vraisemblablement du fait que l'articulation de la consonne latérale /l/ (et, partant, l'impédance du conduit vocal et le débit oral) est fortement influencée par le degré d'arrondissement des voyelles adjacentes.

REMERCIEMENTS

Vifs remerciements à Jacqueline Vaissière, qui a dirigé cette étude, ainsi qu'à Lise Crevier-Buchman et Bernard Roubeau pour la prise des données.

BIBLIOGRAPHIE

[1] A. Amelot. *Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français*, thèse, Univ. Paris 3, 2004.

[2] A. Amelot, B. Roubeau, L. Crevier-Buchman et S. Maeda. Prise de données simultanées aérodynamiques et fibroscopiques durant les voyelles nasales : comparaison avec des données prises séparément. *Actes des XXV[e] JEP*, 2004.

[3] R.J. Baken. *Clinical Measurement of Speech and Voice*. Taylor & Francis, Londres, 1987.

[4] P. Basset, A. Amelot, J. Vaissière et B. Roubeau. Nasal airflow in French spontaneous speech, *J. of the IPA*. 31(1):87-100, 2001.

[5] A.-P. Benguerel. Nasal airflow patterns and velar coarticulation in French. In *Speech Wave Processing and Transmission*, dir. par G. Fant, Almqvist & Wiksell, Stockholm, 105-112, 1974.

[6] V. Delvaux. Etude aérodynamique de la nasalité en français. *Actes des XXIII[e] JEP*, 2000.

[7] V. Delvaux. *Contrôle et connaissance phonétique : Les voyelles nasales du français*, thèse, Univ. Libre de Bruxelles, 2003.

[8] M. Durand. *Le genre grammatical en français parlé à Paris et dans la région parisienne*. D'Artrey, Paris, 1936.

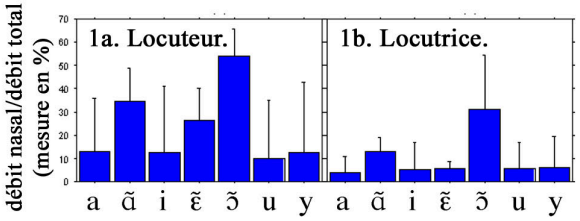
[9] A. Ghio et B. Teston. Caractéristiques de la dynamique d'un pneumotachographe pour l'étude de la production de la parole : aspects acoustique et aérodynamique. *Actes des XXIV[e] JEP*, 2002.

[10] B. Hayes et T. Stivers. A phonetic account of postnasal voicing, ms., Univ. de Californie, 2000.

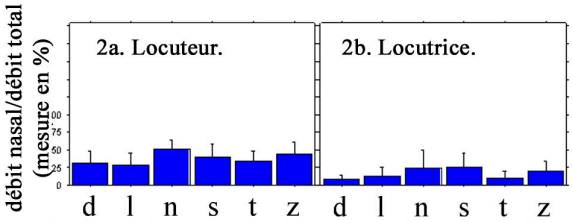
[11] R. Krakow. Nonsegmental influences on velum movement patterns: syllables, sentences, stress

and speaking rate. In *Phonetics and phonology, Volume 5*, dir. par M. Huffman et R. Krakow, Academic Press, San Diego, 87-116, 1993.

[12] J. Ohala. Phonetic explanations for nasal sound patterns. In *Nasálfest: Papers from a symposium on nasals and nasalization*, dir. par C.A. Ferguson *et al.*, Stanford, 289-316, 1975.



Graphique 1. Proportion de débit oral et nasal pendant Vt, en fonction de la voyelle concernée.



Graphique 2. Proportion de débit nasal et oral pendant Vt en fonction du contexte consonantique. *Ces données correspondent uniquement aux trois voyelles nasales.*

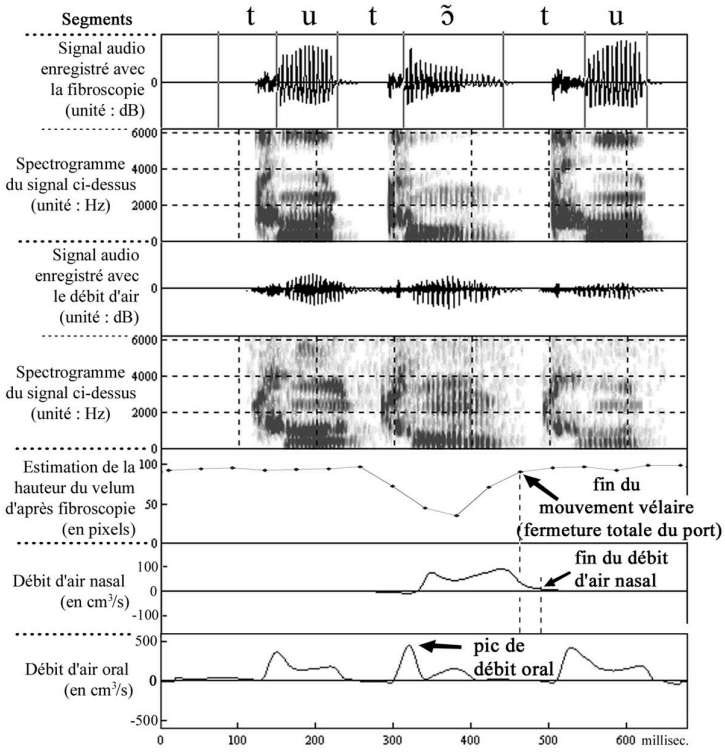


Fig. 1. Exemple d'alignement des données audio, aérodynamiques (débit oral et nasal) et fibroscopiques. Logatome /tutɔ̃tu/, locuteur M.

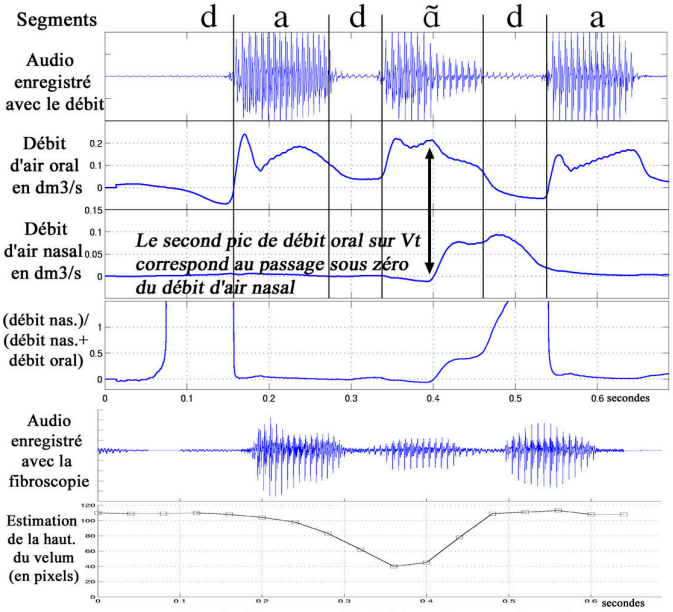
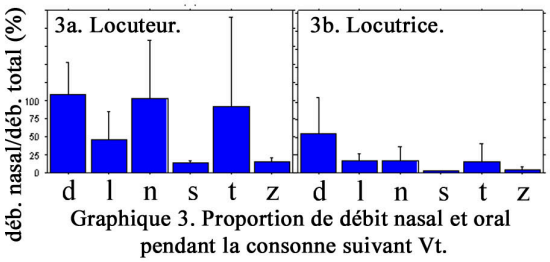
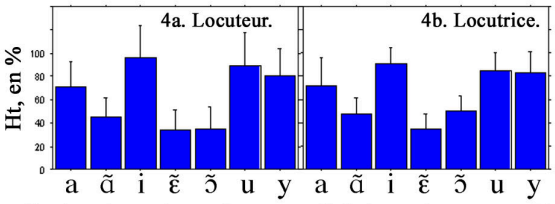


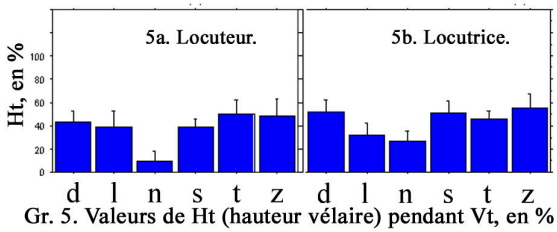
Fig. 2. Signaux audio, débit oral et nasal, et Ht, logatome dadãda, locuteur masculin.



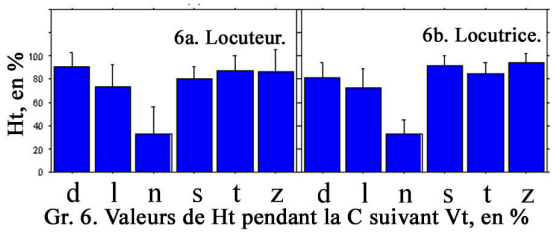
Graphique 3. Proportion de débit nasal et oral pendant la consonne suivant Vt.



Gr. 4. Valeurs de Ht (hauteur vélaire) pendant Vt, en % de la hauteur observée pendant la tenue d'une occlusion.



Gr. 5. Valeurs de Ht (hauteur vélaire) pendant Vt, en % de la hauteur observée pendant la tenue d'une occlusion.



Gr. 6. Valeurs de Ht pendant la C suivant Vt, en % de la hauteur observée pendant la tenue d'une occlusion.

Sensibilité au débit et marquage accentuel des phonèmes en français

Valérie Pasdeloup°, Robert Espesser° & Malika Faraj*

*Université de Rennes 2
°Laboratoire Parole et Langage, UMR 6057 CNRS, Université de Provence, France
valeriepasde@yahoo.fr, robert.espesser@lpl.univ-aix.fr, malika.faraj@free.fr

ABSTRACT

The aim of this work is to determine the way the prosodic scene reorganises itself according to speech rate variations in French. We present the temporal structure study of a one thousand word speech corpus. The corpus was produced at three different rates (normal, fast and slow) by one speaker with two repetitions. The goal is to study the relationship between speech rate sensitivity and accentual markedness of phoneme. Results put in light that phoneme does not behave the same way if stressed or not and if consonantic or vocalic. Unstressed phonemes are less rate sensitive than stressed ones. Vowels are more rate sensitive than consonants, especially when stressed. Nevertheless, consonants are rate sensitive in such proportions when stressed that it is not possible to say, as usually said, that it is the vowel which carries stress.

1. INTRODUCTION

1.1. L'organisation de la scène prosodique

Le décalage entre le percept d'une forme visuelle et le stimulus sur lequel il se base est un phénomène bien connu dans les illusions visuelles étudiées par la théorie de la Gestalt. Ce décalage illustre un des aspects génératifs de la perception. On peut expérimenter un phénomène similaire dans la perception du flux rythmique d'un texte lu à différents débits de parole. Par exemple, à débit lent on a l'impression subjective que l'ensemble du flux rythmique (c'est-à-dire toutes les syllabes) est produit plus lentement qu'à débit normal, alors qu'en fait toutes les syllabes ne sont pas également sensibles aux variations de débit : les syllabes inaccentuées sont moins sensibles au débit que les syllabes accentuées [14, 15].

Cette étude de la scène prosodique se situe dans le cadre de la théorie de la *Gestalt*. Cette théorie qui a beaucoup été appliquée à la perception des sons [8, 9] et à celle de la musique [3, 4, 5] a eu peu d'écho dans l'étude de la prosodie [12, 2, 1, 15]. Une des lois fondamentales de la théorie de la forme est que nous percevons des *figures*, des *formes*, qui se dégagent d'un *fond* en vision et en audition. Ces formes correspondent à des « objets sensibles » qui possèdent des caractéristiques spécifiques afin de pouvoir ainsi émerger d'un fond [10]. Les figures n'ont aucune existence autonome puisqu'elles n'existent qu'en relation avec un fond. La figure a comme caractéristique fonctionnelle de posséder une forme et une organisation alors que le fond est une continuité amorphe, indéfinie qui n'a pas de contours propres. En parole, dans l'optique de la théorie de la Gestalt, les suites de syllabes non-accentuées constituent le fond de la scène prosodique et les syllabes accentuées constituent les figures qui *émergent* de ce fond. Ainsi, une suite de syllabes constitue en tant que suite amorphe et indéfinie le fond de la scène prosodique, tant que n'émerge pas un accent qui donne forme à une syllabe particulière. Il n'y a donc pas à proprement parler de syllabes inaccentuées, mais des syllabes qui reçoivent ou non un accent.

1.2. Objectifs

La plupart des recherches sur l'influence du débit de parole est consacrée à l'étude des unités segmentales. Peu de travaux ont été consacrés en français aux effets du débit de parole sur l'organisation prosodique [6, 7, 18].

Cette recherche prend place dans un projet plus large sur les gabarits rythmiques et la pulsation accentuelle en français. Le but est de contraindre la structuration rythmique de textes lus en manipulant le débit de parole afin d'observer les contraintes qui opèrent sur la production de la pulsation accentuelle et le formatage des gabarits rythmiques. Dans une étude précédente [14, 15], nous avons mis en évidence que les variations de débit n'ont pas la même influence dans la scène prosodique sur les syllabes selon qu'elles jouent le rôle de forme ou de fond : le fond de syllabes inaccentuées est moins élastique que les formes (les syllabes accentuées) qui en émergent. L'objectif du présent travail est d'étudier comment ce phénomène opère à l'intérieur de la syllabe sur la voyelle et la consonne. Les voyelles sont-elles plus sensibles au débit que les consonnes ? Nous présentons ici les résultats relatifs à l'étude de la structuration temporelle des phonèmes d'un corpus lu de mille mots dans trois conditions de débit de parole par une locutrice.

2. MÉTHODOLOGIE EXPÉRIMENTALE

2.1. Corpus

Le corpus est un conte d'environ 1000 mots, lu en chambre sourde, dans trois conditions de débit (normal, rapide et lent), par une locutrice (la 1ère auteure), avec deux répétitions. La meilleure répétition a été ensuite retenue à chaque débit. Les corpus lus correspondent à environ 1200 syllabes pour chaque condition de débit. Pour les trois débits, on totalise 8081 phonèmes : 2660 à débit rapide, 2698 à débit normal et 2723 à débit lent. Parmi ces 8081 phonèmes, on décompte : 4085 consonnes, 3527 voyelles (dont 423 schwas constituant

un noyau vocalique), 108 schwas extra-métriques (ne constituant pas le noyau vocalique d'une syllabe, en général devant pause) et 361 semi-voyelles.

2.2. Analyse expérimentale

En parole, la composante acoustique du rythme correspond selon nous à tout stimulus acoustique qui seul ou en interaction avec d'autres permet de produire un percept rythmique : contrastes mélodiques, de durée, d'intensité et de timbre. L'étude de la structuration rythmique des énoncés du corpus inclut par conséquent celle de la prosodie (intonation et accentuation) : l'analyse phonétique des paramètres prosodiques (F_0, durée syllabique, phonémique et pauses principalement) et leur interprétation phonologique afin de déterminer une structure rythmique abstraite dans le cadre d'un modèle théorique donné. La représentation phonologique correspond à l'accentuation et aux groupements rythmiques. Notre modèle rythmique distingue quatre niveaux prosodiques [12, 13] :

- la syllabe qui constitue l'unité rythmique minimale et qui peut être accentuée ou non-accentuée (les syllabes accentuées sont indiquées en gras et les limites entre les syllabes par des tirets) ;
- le groupe accentuel qui est le groupement rythmique minimal (indiqué par les symboles < >) ; il est constitué d'une syllabe accentuée précédée généralement d'une ou de quelques syllabes non-accentuées et est soumis à des contraintes de taille ;
- le mot rythmique (indiqué par les symboles [ ]) qui est la plus petite structure prosodique qui organise un groupe de sens (petit groupe syntactico-sémantique) [17] ; il est constitué généralement d'un ou de deux groupes accentuels et est soumis à des contraintes de taille ;
- la séquence rythmique (indiquée par les symboles / /) qui est une structure prosodique de niveau hiérarchique supérieur au mot rythmique qui organise une unité discursive ; elle est constituée en général de plusieurs mots rythmiques mais ne semble pas soumise à des contraintes de taille.

(1) Le rythme$_{2syll}$ d'la parole$_{3syll}$ n'est pas$_{2syll}$ élastique$_{3syll}$ =>

/ [<lə - **ritm**> <dla - pa - **rəl**>] [<ne - **pa**> <e - las - **stik**>] /

(2) Le rhinocéros$_{5syll}$ de Constantinople$_{5syll}$ n'est pas$_{2syll}$ élastique$_{3syll}$ =>

/ [<lə - **ri**> <no - se - **rɔs**>] [<də -**kɔ̃**> <stɑ - ti - **nɔpl**>]
[<ne - **pa**> <e - las - **stik**>] /

Dans notre modèle, les règles phonologiques d'accentuation et d'intonation sont basées sur des contraintes linguistiques au sens strict (morpho-syntaxiques et lexicales) et rythmiques (nombre de syllabes des unités lexicales, des constituants morpho-syntaxiques, des groupes accentuels et des mots rythmiques). Ainsi les énoncés (1) et (2) ci-dessus qui ont la même structure syntaxique mais dont les constituants syntaxiques sont composés d'un nombre différent de syllabes n'auront pas nécessairement la même structure prosodique (pour plus de détails cf. [13]) :

Dans un premier temps, l'étiquetage phonétique des énoncés et leur segmentation phonémique sont effectués avec le logiciel développé au LORIA (D. Foher & Y. Laprie : http://www.loria.fr/equipes/parole/) puis corrigés manuellement. Le logiciel code identiquement les voyelles orales et nasales à double timbre correspondant aux archiphonèmes : /E Œ O A E~/. On obtient ainsi 11 types de consonnes et 17 types de voyelles, en excluant les schwas extra-métriques et les semi-voyelles qui n'ont pas été pris en compte dans l'analyse des phonèmes. La syllabation est effectuée avec un script sous Praat et corrigée manuellement. Dans un second temps, l'analyse phonétique des paramètres prosodiques est réalisée avec l'appui de l'écoute perceptive par la 1[ère] et la 3[ème] auteures. Le caractère accentué ou non accentué des syllabes est ainsi déterminé en fonction des contrastes accentuels effectivement réalisés et perçus. Dans un troisième temps, ces données sont interprétées dans le cadre de notre modèle prosodique ce qui permet de préciser le type d'accent (primaire ou secondaire) et le type de groupement rythmique. Dans la présente étude, seule l'interprétation de l'accent, c'est-à-dire le statut accentué ou inaccentué des phonèmes, a été pris en compte. Pour une étude de l'influence du débit sur les pauses et les syllabes dans ce corpus cf. [15].

3. RÉSULTATS

3.1. Taux d'articulation et durée des phonèmes

Le taux d'articulation est de 15.31 phonèmes/s à débit rapide (R), 12.33 phon/s à débit normal (N) et 9.88 phon/s à débit lent (L) (hors pauses ; schwas extra-métriques et semi-voyelles inclus). Pour les syllabes, le taux d'articulation est de 6.8 syll/s à débit R, 5.4 syllabes à débit N et 4.4 syll/s à débit L.

Table 1 : Moyenne des durées syllabiques des Consonnes et Voyelles Inaccentuées et Acc. dans les trois débits

	débit Rapide	débit Normal	débit Lent	moyenne
C I	62.06ms ±25	68.41ms ±28	79.57ms ±33	70.01
C A	79.06ms ±34	95.35ms ±41	119.06ms ±51	97.82
V I	57.83ms ±18	68.08ms ±21	80.55ms ±26	68.82
V A	68.99ms ±24	104.21ms ±39	144.62ms ±75	105.94

La Table 1 fait apparaître que les phonèmes accentuées (A) sont plus sensibles au débit que les phonèmes inaccentuées (I). Ce phénomène semble plus manifeste chez les voyelles (V) accentuées qui sont plus sensibles au débit que les consonnes (C) accentuées : comparée au débit N, la durée des VA varie en moyenne de –34% à débit R et de +39% à débit L, alors que celle des CA varie en moyenne de –17% à débit R et de + 25% à débit L. Les CI et VI manifestent une insensibilité au débit très proche

: 18ms de différence moyenne entre les deux conditions extrêmes de débit R et L pour les CI contre 23ms pour les VI. La plus grosse différence de durée entre C et V s'observe chez les accentuées à débit lent.

3.2. Traitement statistique

La durée des phonèmes a été évaluée en fonction de trois facteurs : *Débit*, facteur ordonné à 3 niveaux (Rapide, Normal, Lent) ; *Accent*, facteur à 2 niveaux (Inaccentué, Accentué) ; *CLasse*, facteur à 2 niveaux (Consonne, Voyelle). Un modèle linéaire mixte, où le phonème est le facteur de groupement, a permis de traiter la répétition des 28 groupes non équilibrés de phonèmes (17 consonnes et 11 voyelles) ([16], http://www.R-project.org/). Ainsi, les variations inter-phonémiques de durée ont été neutralisées. De plus, l'utilisation du logarithme de la durée a stabilisé la variance. Ce premier modèle ayant montré que seules les composantes linéaires du débit sont significatives, le facteur débit a été considéré ensuite comme une variable numérique classique, ce qui simplifie le modèle. A chaque débit a été associée la durée totale correspondante du corpus lu (hors pauses) : 175s pour le débit R, 225s pour le débit N et 275s pour le débit L. La variable *Débit* a été centrée sur le débit R afin de tester certaines hypothèses spécifiques à ce débit.

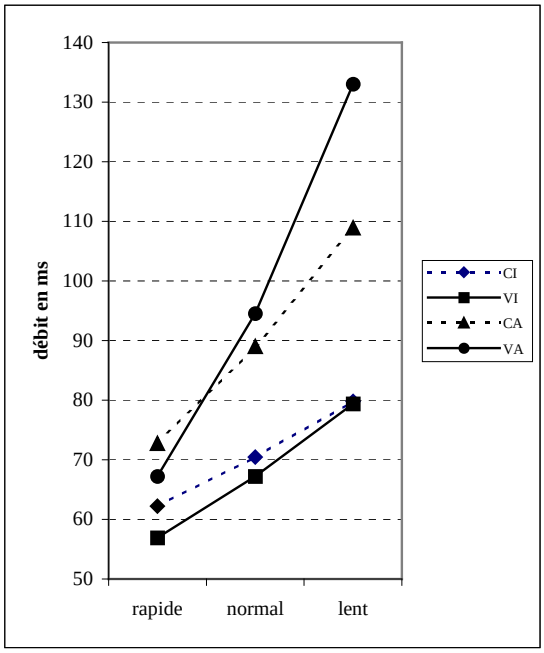


Figure 1 : Durées estimées des consonnes et des voyelles par le modèle mixte dans les trois conditions de débit

La table 2 des régresseurs du modèle montre que tous les coefficients d'interaction avec le débit sont significatifs et positifs : *Débit:AccentA*, *Débit:CLasseV*, *Débit: AccentA:ClasseV*. On a donc quatre droites de régression distinctes CI, VI, CA et VA (Fig. 1). La significativité des autres coefficients *AccentA* (significatif), *ClasseV* (non significatif) et *AccentA:ClasseV* (non significatif) n'a de valeur que pour le débit R.

Table 2 : Régresseurs du modèle mixte centré

	Value	Std.Error	DF	t-value	p-value
(Intercept)	4.1308	0.0612	7577	67.47	0.0000
Débit	0.0025	0.0002	7577	14.54	0.0000
AccentA	0.1556	0.0164	7577	9.48	0.0000
ClasseV	-0.0894	0.0973	26	-0.92	0.3667
Débit:AccentA	0.0016	0.0002	7577	6.27	0.0000
Débit:ClasseV	0.0008	0.0002	7577	3.47	0.0005
AccentA: ClasseV	0.0108	0.0244	7577	0.44	0.6583
Débit:AccentA: ClasseV	0.0019	0.0004	7577	5.30	0.0000

3.3. Interprétation

L'interaction significative *Débit:AccentA* montre que les phonèmes inaccentués sont moins sensibles au débit que les phonèmes accentués (cf. Table 1 et Fig. 1). Ces résultats sont comparables à ceux trouvés par Duez [6]. En français, les variations de débit n'ont donc pas le même effet sur les phonèmes selon qu'ils jouent le rôle de fond ou de forme dans la scène prosodique. L'interaction *Débit:CLasseV* montre que l'effet du débit est plus marqué pour les V que pour les C. La moins grande élasticité des consonnes a été observée dans d'autres travaux [6, 11]. La double interaction *Débit:AccentA: CLasseV* précise que l'accentuation renforce la sensibilité au débit des V. Par conséquent, les V sont plus sensibles au débit que les C, et ce surtout chez les accentuées et de façon très peu marquée mais significative chez les inaccentuées.

Selon notre hypothèse, cette différence de sensibilité au débit des C et des V serait liée à des *contraintes universelles de matière* (contraintes motrices de contrôle articulatoire et contraintes sensori-motrices proprioceptives et auditives). En effet, du fait que les consonnes sont par nature des phénomènes transitoires, elles ne peuvent s'allonger au delà d'une certaine limite quand le débit ralentit (effet plafond). En revanche, la différence de sensibilité au débit des phonèmes accentués et inaccentués serait phonologique, c'est-à-dire liée à des *contraintes de forme* puisqu'elle résulterait de la structuration formelle de la scène prosodique.

Par ailleurs, la plus faible sensibilité des C au débit a pour conséquence qu'à débit R les V ne se distinguent plus des C à la fois chez les inaccentuées et chez les accentuées (*AccentA:ClasseV* : non significatif). En extrapolant à débit encore plus rapide, on peut émettre l'hypothèse que la distinction entre les C et les V deviendrait significative, les C devenant alors plus longues que les V.

Enfin, du point de vue des contrastes accentuels de durée entre les phonèmes inaccentués et accentués (écart A-I), les contrastes se renforcent quand le débit ralentit et sont significativement plus marqués chez les V que chez les C à débit N et L. A débit R, le contraste accentuel subsiste (*AccentA* : significatif) mais ne diffère plus significativement entre les V et les C (*AccentA:ClasseV* : non significatif). En extrapolant à débit encore plus rapide, on peut supposer que ce contraste deviendrait

significatif, le contraste accentuel de durée chez les C devenant plus marqué que chez les V.

4. DISCUSSION

Dans cette étude, nous avons montré qu'en français la consonne tout comme la voyelle est plus sensible au débit quand elle est accentuée qu'inaccentuée. Ce phénomène de plus grande sensibilité au débit, que nous avions déjà observé pour la syllabe accentuée [14, 15], opère donc sur les deux constituants syllabiques. La consonne est cependant moins sensible au débit que la voyelle, et ce surtout chez les accentuées. En ce qui concerne la programmation motrice, nous émettons l'hypothèse que seule la durée des phonèmes accentués serait planifiée (plus précisément le contraste temporel). La durée des phonèmes non-accentués ne serait pas planifiée. Par conséquent, la grande sensibilité des phonèmes accentués au débit correspondrait à des variations de haut niveau dans le système (commandes motrices), alors que la très faible sensibilité des phonèmes inaccentués correspondrait à des variations de bas niveau.

Dans la scène prosodique, les objets consonnes et voyelles - bien que différemment contraints dans leur matérialité - seraient phonologiquement identiques pour constituer soit des formes en émergeant, soit le fond en restant amorphes et indéfinis. Les consonnes et les voyelles inaccentuées, du fait de leur relative insensibilité au débit, participeraient conjointement à l'illusion perceptive du débit : l'auditeur a l'impression subjective que le fond de la scène prosodique accélère ou décélère selon les différents débits, alors qu'en fait c'est principalement la durée de présentation des formes qui se modifie sur un fond relativement stable. Quand le débit s'accélère, la durée de présentation des phonèmes accentués diminue. Quand le débit ralentit, la durée de présentation des phonèmes accentués s'allonge.

En conclusion, sur le plan de la substance, ce n'est pas uniquement la voyelle qui "porte l'accent" dans la syllabe. La consonne a une contribution non négligeable dans la réalisation du contraste temporel entre les syllabes inaccentuées et accentuées. Enfin, en ce qui concerne la relation entre le marquage phonologique suprasegmental et la variation phonétique, on observe que les phonèmes marqués par l'accent sont plus sensibles à la variation que les phonèmes non-marqués.

REMERCIEMENT : Nous souhaitons remercier Daniel Hirst pour ses conseils ainsi que pour la réalisation de nombreux scripts sur Praat.

BIBLIOGRAPHIE

[1] C. Astésano. *Rythme et Accentuation en Français : Invariance et Variabilité stylistique*, Paris, L'Harmattan, 2001.

[2] E. Couper-Kuhlen. *English Speech Rhythm: Form and function in everyday verbal interaction, Amsterdam*, John Benjamins Publishing Compagny, 1993.

[3] D. Deutsch. Grouping mechanism in music, in Deutsch D. (ed.), *The psychology of music*, New York, Academic Press, 99-130, 1982.

[4] C. Drake. *Processus cognitifs impliqués dans l'organisation du rythme musical*, Thèse doctorale, Université René-Descartes, Paris, 1990.

[5] C. Drake. Reproduction of musical rhythms by children, adult musicians and adult nonmusicians, *Perception & Psychophysics*, 53 (1), 25-33, 1993.

[6] D. Duez. *Contribution à l'étude de la structuration temporelle de la parole en français*, Thèse de Doctorat d'Etat, Université de Provence, Aix-Marseille 1, 1987.

[7] C. Fougeron & S.-A. Jun. Rate effects on French intonation: prosodic organization and phonetic realization, *Journal of Phonetics*, 26, 45-69, Academic Press, 1998.

[8] P. Fraisse. Les structures rythmiques, *Studia Psychologica*, Publications Universitaires de Louvain, 124p.,1956.

[9] P. Fraisse. *Psychologie du rythme*, Paris, Presses Universitaires de France, 360p., 1974.

[10] P. Guillaume *La psychologie de la forme*, (1937), Paris, Flammarion, 1979.

[11] A. Kozhevnikov & L. A. Chistovich C. Speech articulation and perception, in *Joint Publications Research Service*, 543p., 1965.

[12] V. Pasdeloup. *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*, Thèse de doctorat, Université de Provence Aix-Marseille 1, 1990.

[13] V. Pasdeloup. A prosodic model for French text-to-speech synthesis: A psycholinguistic approach, in Bailly G., Benoît C. & Sawallis T.R. (eds.), *Talking Machines: Theories, Models and Designs*, Elsevier Science Publisher, 335-348, 1992.

[14] V. Pasdeloup. Figures et fond dans la scène prosodique : leur résistance face aux variations du débit de parole, *1[er] Symposium International « Interface Discours-Prosodie »*, 8-9 sept. 2005.

[15] V. Pasdeloup, R. Espesser & M. Faraj. Rate sensitivity of syllable in French: a perceptual illusion? *3[rd] International Conference on « Speech Prosody »* , Dresde, 2-5 mai 2006

[16] J. C. Pinhero & D. M. Bates. *Mixed-Effects Models in S and S-Plus*, Springer, 2001.

[17] J. Vaissière. "La structuration acoustique de la phrase française", *Annali della Scuola Normale Superiore di Pisa*, 3(10), 529-560, 1980.

[18] B. Zellner. *Caractérisation et prédiction du débit de parole en français*, Doctorat, Lausanne, 1998.

Différenciation des mots de fonction et des mots de contenu par la prosodie : analyse d'un corpus trilingue de langage adressé à l'enfant et à l'adulte

Christelle Dodane, Jean-Marc Blanc** & Peter Ford Dominey***

*Laboratoire Dynamique du Langage, UMR CNRS 5596
14 avenue Berthelot, 69 363 LYON Cedex 07 - France
dodane@isc.cnrs.fr
**Institut des Sciences Cognitives, UMR CNRS 5015
67 Bd Pinel, 69675 BRON Cedex – France
dominey@isc.cnrs.fr ; blanc@isc.cnrs.fr

ABSTRACT

This research investigated the role of salient prosodic cues in the first approximant assignment in function and content words by infants. In order to discoverer these cues and to see if they could vary cross-linguistically, infant-directed speech was compared to adult-directed speech in three different languages, French, English and Japanese. The same story was successively read by 15 mothers to their infant and to an adult (5 mothers per language). The acoustic analyses reveal that non-final syllable duration, Fo peaks and relative amplitude and amplitude peaks are particularly relevant in the three languages to allow categorization in function and content items, but they have a different relative weight across languages because of the specific prosodic organization of these languages.

1. INTRODUCTION

Afin de pouvoir commencer à élaborer le savoir syntaxique de sa langue maternelle, le bébé doit d'abord être capable de segmenter le signal de parole en constituants. Or, comment l'enfant peut-il commencer à segmenter le flux de parole alors que les différents paramètres acoustiques varient de façon continue et qu'il existe une très grande variabilité intra et interlocuteurs ? Les théories d'initialisation prosodique supposent que le signal de parole contient des signaux prosodiques réguliers et suffisamment saillants pour permettre à l'enfant d'accéder à une information grammaticale de type rudimentaire sur les principales catégories linguistiques de sa langue. La distinction entre mots de fonction et mots de contenu pourrait notamment guider l'enfant sur la route de l'initialisation syntaxique (Gleitman, [1]) et sémantique (Pinker [2]). Une segmentation de ce type serait très utile car, bien que les mots de fonction soient en nombre restreint, ils ont une très grande fréquence dans la parole. Leur repérage permettrait d'accéder à un grand nombre de frontières de mots (Christophe & al., [3] ; Ramus [4]). Il se trouve que les mots de fonction sont minimalisés dans leur forme parlée (Culter [5] ; Shi [6] ; Selkirk [7] ; Morgan & al. [8] ; Shi & al. [9]), et en particulier, au niveau prosodique (Selkirk [7]). Les deux classes de mots ont donc un poids perceptuel très différent, les mots de contenu étant prosodiquement plus saillants. Les bébés sont sensibles à ces différences puisque dès la naissance, ils sont capables de discriminer les mots de contenu des mots de fonction (Shi & al. [10]) et à 6 mois, ils marquent une préférence pour les mots de contenu (Shi & Werker [11]) et se focalisent sur les mots ayant un fort poids sémantique (acquisition du lexique). La sensibilité aux mots de fonction apparaît plus tardivement, entre 8 et 13 mois (Shi & al. [12]). Si cette distinction entre mots de contenu et mots de fonction est importante pour l'enfant, il est probable qu'elle soit exagérée par les parents lorsqu'ils s'adressent à leurs enfants (langage adressé à l'enfant, désormais L.A.E.) dans un processus de « pédagogie inconsciente ». Le L.A.E. étant une modalité de langage où le niveau prosodique est largement exagéré (Papousek & Papousek [13] ; Fernald [14]), il est probable que les indices prosodiques pertinents pour une telle distinction soient également exagérés. Afin de déterminer quels indices prosodiques pourraient permettre aux bébés de différencier les mots de contenu des mots de fonction, il serait donc intéressant de comparer deux modalités de langage, le LAE et le langage adressé à l'adulte (désormais LAA). Par ailleurs, ces indices devraient diverger en fonction des caractéristiques du système prosodique de la langue étudiée. Nous proposons donc de comparer deux modalités de langage (LAA vs LAE), et ce, dans trois langues appartenant à des catégories prosodiques différentes, l'anglais (tendance à la rythmicité accentuelle), le français (tendance à la rythmicité syllabique) et le japonais (langue à rythmicité moraïque). Nous établirons un relevé contrastif permettant d'évaluer le rôle de chaque indice au sein 1°) de chaque langue 2°) de toutes les langues étudiées 3°) en fonction des deux modalités étudiées.

2. MÉTHODOLOGIE

Participants

15 mères volontaires ont participé à cette étude, 5 mères françaises, 5 mères japonaises et 5 mères anglaises. Si toutes les mères ont été enregistrées à Lyon, les mères japonaises et les mères anglaises ne s'adressaient à leur enfant que dans leur langue maternelle. L'âge de leur enfant allait de 6 mois à 22 mois.

Enregistrements

Les mères ont toutes été enregistrées à leur domicile, par le père ou une personne familière afin d'obtenir les enregistrements les plus naturels possibles. Il était demandé aux mères de lire la même histoire à un adulte familier (modalité LAA), puis à leur enfant (modalité LAE). Les enregistrements étaient dissociés dans le temps afin que les modalités soient nettement différenciées. L'histoire était extraite d'un livre pour enfant français (Cousins [15]). Elle a été traduite en anglais et en japonais par des locuteurs natifs. Les enregistrements ont été réalisés avec un enregistreur minidisque (Sony MZN-910S) et un micro unidirectionnel Philips SBC-MD695. Chaque mot, chaque syllabe et chaque phonème a été étiqueté en fonction de leur appartenance à une catégorie grammaticale générique (fonction vs contenu) et une catégorie grammaticale spécifique (nom, verbe, adverbe, adjectif, déterminant, auxiliaire, conjonction, pronom et enclitique uniquement pour le japonais). Pour le japonais, le codage a été vérifié par un linguiste et locuteur natif du japonais.

Analyses acoustiques

Les phrases ont été échantillonnées à 22 kHz, 16 bits en mono. Les analyses acoustiques ont été réalisées avec le logiciel Praat. Le contour de Fo a été extrait, puis post-traité (suppression des sauts d'octave, lissage, interpolation). Un prosogramme a été édité (Mertens [16]) à partir du contour de Fo. Tous les événements mélodiques affichant une montée suivie d'une descente sont considérés comme des accents de hauteur (pics de Fo). La totalité des accents relevés grâce à cette méthode ont été vérifiés pour les trois langues par un phonéticien et musicien entraîné. La durée de la totalité des mots, des syllabes et des phonèmes (en ms) a été extraite automatiquement à partir des grilles de segmentation de Praat. Le contour d'amplitude a été extrait avec Praat et l'amplitude relative de chaque syllabe a été déterminée en calculant le rapport entre son énergie (RMS, root-mean square en dB) et l'énergie de la syllabe la plus intense de l'énoncé (Shi & al. [9]).

	Mots				Syllabes			
	F type	F tok.	C type	C tok.	F type	F tok.	C type	C tok.
Anglais	19	42	44	73	23	44	59	107
Français	21	44	42	66	22	33	71	128
Japonais	17	48	42	68	22	52	67	149

Table 1 : Nombres de types et de token (abrégé en tok.), de mots et de syllabes pour les mots de contenu (C) et les mots de fonction (F) en anglais, français et japonais.

Ces mesures ont permis l'extraction de 33 indices acoustiques différents, construits à partir des mesures de durée (mots, syllabes finales et non finales, voyelles, consonnes, durée moyenne des syllabes portant les pics de Fo), de Fo (Fo moyen, minimum et maximum en Hz, intervalles en ½ tons et en ½ tons par seconde sur les syllabes finales et non finales et les mots - Fo moyen des pics de Fo, relevé des pics de Fo) et d'amplitude (amplitude relative, amplitude moyenne, minimum et maximum en dB, déviation standard sur les syllabes finales et non finales et les mots - relevé des pics d'amplitude).

Analyses statistiques

Des analyses de variance à deux facteurs et à trois facteurs (ANOVA à mesures répétées) ont été réalisées dans le but de déterminer l'influence de la modalité (1[er] facteur : LAA - LAE), de la catégorie lexicale (2[ème] facteur : mot de contenu - mot de fonction) et de la langue (3[ème] facteur : français - anglais - japonais).

3. RÉSULTATS

3.1. Résultats et discussion

Résultats généraux

Parmi les 33 indices prosodiques étudiés, 26 se montrent significativement pertinents (effet de catégorie) pour différencier les items de contenu (désormais C) des items de fonction (désormais F). Les résultats convergent dans les trois langues pour montrer que les C sont prosodiquement plus saillants que les F. Cependant, certains indices jouent un rôle plus important dans cette différenciation. C'est notamment le cas des indices de durée, des pics de Fo, des pics d'amplitude et de l'amplitude relative. Pour cette raison, nous nous limiterons à la présentation de ces indices. Par ailleurs, le rôle respectif de ces indices varie en fonction de la langue et notamment, de son organisation prosodique. Enfin, on relève un effet de la modalité (LAA vs LAE) dans les trois langues, même si cet effet diffère là aussi, en fonction de la langue.

	Anglais	Français	Japonais
Durée des mots	p<,0000*	p<,0002*	p<,0000*
Durée SNF	p<,0004*	p<,0030*	p<,0811
Pics de Fo	p<,0003*	p<,0000*	p<,0002*
Amplitude relative	p<,0020*	p<,0003*	p<,0018*
Pics d'amplitude	p<,0002*	p<,0002*	p<,0000*

Table 2 : Significativité des indices les plus saillants dans les trois langues étudiées en fonction de la catégorie (F/C)

Durée des syllabes

En anglais et en français, la durée des syllabes appartenant aux mots de C est significativement plus grande que celles appartenant aux mots de F. Ainsi, en anglais, on relève une augmentation de 127 ms de F à C en LAA et de 154 ms de F à C en LAE (effet de catégorie : p<.0000* ; effet de modalité : p<.0032*). En français, on relève également une augmentation de durée, mais bien moins marquée qu'en anglais avec une augmentation de 44 ms de F à C en LAA et de 62 ms de F à C en LAE (effet de catégorie : p<.0009* ; effet de modalité : p<.0142*). En revanche, pour le japonais, les syllabes de F sont plus longues que

les syllabes de C ; en effet, on relève une baisse de 15 ms de F à C en LAA et de 39 ms de F à C en LAE (pas d'effet de catégorie). Ces résultats s'expliquent par le fait que la majorité des énoncés japonais dans notre corpus se termine par une particule enclitique, c'est-à-dire un mot de F. Par ailleurs, l'allongement final (AF) est spécialement marqué en LAE pour les trois langues de notre corpus. Ainsi, en anglais, la durée moyenne des syllabes finales (SF : syllabes précédant une pause) est de 484 ms en LAA et de 619 ms en LAE, soit une augmentation de 132 ms d'une modalité à l'autre. En français, on relève une durée moyenne de 312 ms en LAA et de 442 ms en LAE, soit une augmentation de 130 ms et en japonais, une durée moyenne de 272 ms en LAA et de 398 ms en LAE, soit une augmentation de 126 ms. Du point de vue perceptif, l'AF est particulièrement utile pour le bébé car il lui permet de segmenter le signal de parole en constituants de grande taille (Koponen & Lacerda [17], les propositions et les phrases. Cependant, les différences de durée sont également importantes pour repérer des constituants plus petits, tels que les mots F ou de C. L'AF venant perturber la mesure de la durée des syllabes, nous avons choisi de ne conserver que les syllabes non finales (SNF), c'est-à-dire celles qui ne portent pas l'AF (figure n°1).

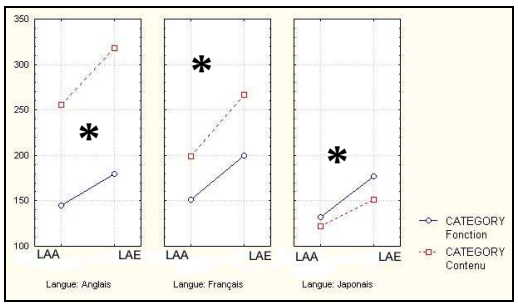


Figure 1 : Durée des finales non finales (en ms) des mots de F et des mots de C en LAA et LAE avec de gauche à droite, l'anglais, le français et le japonais.

Ainsi, dans les trois langues, les SNF des C sont significativement plus longues que dans les F. Cette différence est beaucoup plus marquée en anglais (augmentation de 63 ms en LAA et de 77 ms en LAE), qu'en français (augmentation de 33 ms en LAA et de 25 ms en LAE) et en japonais (augmentation de 15 ms en LAA et de 19 ms en LAE). Cette grande différence de durée en anglais s'explique par le fait que les accents sont localisés sur les mots de C. Tandis que les syllabes portant l'accent sont dilatées au niveau temporel, les syllabes F, qui sont atones, sont compressées. Si l'on pouvait prévoir ce type de résultat en anglais en raison de la rythmicité accentuelle de cette langue, les résultats sont plus surprenants en français. En effet, bien que le français soit décrit comme une langue dont les syllabes non finales ont une tendance à l'isochronie, les SNF des C sont plus longues (+ 33 ms en LAA et + 25 ms en LAE). En ce qui concerne l'effet de la modalité, il est significatif en français (p<.0173*) et en anglais (p<.0055*), mais pas en japonais (p<.0786).

Pics de Fo

Les mots de C sont mis en relief par la présence très fréquente d'une montée intonative ou d'un pic de Fo (qu'on relève en moyenne dans plus de 60% des mots de C en position non finale). Cet indice se révèle hautement significatif dans les trois langues (voir figure n°2) et il semble particulièrement important en français (p<.0000*) et en japonais (p<.0002*). Pour le français, ces résultats rejoignent ceux de Blanc & al. [18] qui montrent qu'un réseau de neurones est capable de reconnaître 83,6 % (corpus LSCP) et 70,3 % (corpus Multext) des mots de C à partir du seul contour de Fo. En revanche, on ne relève aucun effet de modalité dans les trois langues. La proportion de pics de Fo est la même en LAA et en LAE, leur répartition dépendant de la structure prosodique et linguistique des énoncés. S'il y a bien un effet de modalité, il concerne la hauteur des pics de Fo. Ainsi, les pics sont significativement plus élevés en LAE qu'en LAA en français (337 Hz en LAA ; 481 Hz en LAE) et en anglais (324 Hz en LAA ; 400 Hz en LAE). En revanche, l'effet de modalité n'est pas significatif en japonais (225 Hz en LAA ; 284 Hz en LAE).

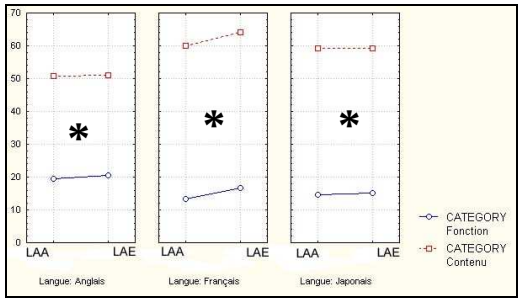


Figure 2 : Proportion (en %) de pics de Fo correspondant aux mots de Cet aux mots de F en LAA et LAE, avec de gauche à droite, l'anglais, le français et le japonais.

Amplitude relative et pics d'amplitude

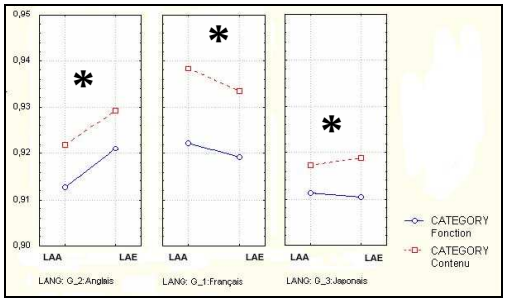


Figure 3 : Amplitude relative des mots de C et des mots de F en LAA et LAE, avec de gauche à droite, l'anglais, le français et le japonais.

L'amplitude relative des SNF est significativement plus élevée sur les C que sur les F en anglais, en japonais et spécialement en français. Ces résultats vont dans le même sens que ceux de Shi & al. [9] pour le Mandarin et le Turc. En revanche, l'effet de la modalité est différent en fonction des langues (figure n°3). En français, l'amplitude décroît pour les F et les C en passant du LAA au LAE. En

anglais, c'est le contraire puisque l'amplitude croît pour les F et les C en passant du LAA au LAE. Enfin, en japonais, il n'y a pas de différence marquée entre le LAA et le LAE. Par ailleurs, la majorité des pics d'amplitude sont localisés sur les mots de C. En effet, plus de 70% des pics sont localisés sur les mots de C dans les trois langues, contre moins de 30% sur les F. Cet indice est spécialement marqué en japonais et en français.

4. CONCLUSION

Les indices étudiés révèlent que les mots de C sont prosodiquement beaucoup plus saillants que les mots de F en français, en anglais et en japonais. Ainsi, la durée des SNF est plus longue pour les C, en particulier pour l'anglais. Les mots de C sont mis en relief par la présente très fréquente de pics de Fo, spécialement en français et en japonais. Enfin, l'amplitude relative est plus élevée sur les syllabes de C et les pics d'amplitude sont le plus souvent localisés sur les mots de C. Dans notre corpus, la « minimalisation » acoustique des F relevée en français, en anglais et en japonais rejoint les résultats de Shi & al. [9] pour l'anglais, le turc et le mandarin. Cependant, il serait nécessaire de compléter cette étude par une analyse fine et détaillée des phénomènes accentuels tels que l'accent secondaire en français et des phénomènes d'alignement tonal dans les trois langues. Il serait également intéressant de tester chacun de ces indices avec un réseau de neurones, comme nous l'avons déjà fait pour les pics de Fo [18], afin de pouvoir hiérarchiser leur importance au sein de chaque langue, ainsi que pour toutes les langues étudiées. Par ailleurs, nos résultats font apparaître une augmentation des paramètres prosodiques lorsque l'on passe de la modalité LAA à la modalité LAE et ce, dans les trois langues étudiées. Cependant, tandis qu'on observe une exagération prosodique beaucoup plus forte en anglais, le LAE est beaucoup plus « modéré » en japonais, le français se situant dans une position intermédiaire. Ces variations trouvent sans doute leur origine dans des différences culturelles liées à l'utilisation du LAE dans les langues étudiées.

Nous tenons à remercier Reiko Vacheret pour son aide précieuse concernant l'analyse du japonais, ainsi que toutes les mamans et les bébés qui ont participé à notre étude. Cette étude a été financée par le programme « Human Frontier Science Program ».

BIBLIOGRAPHIE

[1] L. Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1, 3-55, 1990.

[2] S. Pinker. Language learnability and language development. Cambridge: Harvard Univ. Press, 1984.

[3] A. Christophe, T. Guasti, M. Nespor, E. Dupoux, and B. Van Ooyen. Reflections on phonological bootstrapping: its role for lexical and syntactic acquisition. *Language and Cognitive Processes*, 12, (8/6), 585-612, 1997.

[4] F. Ramus. Rythmes des langues et Acquisition du Langage. Thèse de Doctorat, Paris, EHESS, 1999.

[5] A. Cutler. Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistics Research*, 22, 109-131, 1993.

[6] R. Shi. Perceptual correlates of content words and function words in early language input. PhD. Brown University, Providence, 1995.

[7] E. Selkirk. The prosodic structure of function words. In Morgan & Demuth, Signal to Syntax, 187-213, 1996.

[8] J. Morgan, R. Shi & P. Allopenna, P. Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of boostrapping. In Morgan & Demuth, Signal to Syntax, 263-283, 1996.

[9] R. Shi, J. Morgan & P. Allopenna. Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language*, 25, 169-201, 1998.

[10] R. Shi, J. Weker & J. Morgan. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72, B11-B21, 1999.

[11] R. Shi & J. Werker. Six-month-old infants' preference for lexical words. *Psychological Science*, 12/1, 70-75, 2001.

[12] R. Shi, J. Werker & A. Cutler, A. Function words in early speech perception. Proc. of the XVth ICPhs, Barcelona, 3309-3012, 2003.

[13] M. Papousek & H. Papousek. Musical elements in the infant's vocalizations: Their significance for communication, cognition and creativity. In Lipsitt & Rovee-Collier, 163-224, 1981.

[14] A. Fernald. Intonation and communication intent in mother's speech to infants: is the melody the message? *Child Development*, 60, 1497-1510, 1989.

[15] L. Cousins. En route avec Mimi. Albin Michel, 2001.

[16] P. Mertens. Un outil pour la transcription de la prosodie dans les corpus oraux. Traitement Automatique des Langues, 45(2), 109-130, 2004.

[17] E. Koponen & F. Lacerda. Final lengthening in infant-directed speech may function as a cue to phrase constituents. *Phonum*, 9, 9-12, 2003.

[18] Blanc, J.M., Dodane, C. & Dominey, P.F. Temporal Processing for Syntax Acquisition: A simulation Study. Proc. XXVth Ann. Meet. Cognitive Science Society, 2003.

Comment les attitudes prosodiques sont parfois de « faux-amis » : les affects sociaux du japonais *vs.* français

Takaaki Shochi, Véronique Aubergé, Albert Rilliard

Institut de la Communication Parlée, UMR CNRS 5009, Grenoble, France
{Takaaki.Shochi, Veronique.Auberge, Albert.Rilliard}@icp.inpg.fr

ABSTRACT

The attitudes of the speaker during a verbal interaction are affects linked to the speaker's intentions, and are built by the language and the culture. Attitudes are the main part of the affects expressed during everyday interactions. This paper describes several experiments underlying that some attitudes belong both to Japanese and French languages, and are implemented in perceptively similar prosodies, but that some Japanese attitudes don't exist and/or are wrongly decoded by French listeners. Results are presented for the 12 attitudes and three levels of language learning (naive, beginner, intermediary). It has to be noted that French listeners, naive in Japanese, can very well recognize admiration, authority and irritation; that they don't discriminate Japanese question from declaration before the intermediary level, and that the extreme Japanese politeness is interpreted as impoliteness by French listeners, even when they can speak a good level of Japanese.

1. INTRODUCTION

Les affects de la parole sont exprimés par différents niveaux de traitement cognitifs, depuis les expressions contrôlées involontairement (innées) jusqu'aux expressions contrôlées intentionnellement (simulation d'expressions innées ou acquises, dont font partie les attitudes), volontaires de la part du locuteur. Les attitudes et les émotions sont parfois confondues dans la littérature pour certaines de leurs valeurs. Certains affects comme la surprise entrent dans l'une ou l'autre de ces deux catégories, selon les auteurs. Notre point de vue est que la surprise est une émotion quand elle fait partie d'un processus involontaire, mais une attitude quand elle est exprimée volontairement dans la communication. Nous attendons des formes prosodiques qu'elles soient comparables mais gérées dans des temps différents. Lorsqu'un locuteur ne produit aucune attitude lorsqu'il s'exprime, il s'agit d'une attitude « nulle » qui consiste à n'afficher aucune attitude particulière sur son discours.

Ce travail se situe dans une approche interculturelle, dont le but est de recueillir des indices sur d'éventuels universaux à propos des valeurs exprimées et surtout pour les formes prosodiques des expressions. Ce travail porte sur la perception du japonais par les français, les deux langues et les deux cultures sont éloignées et comportent nombre de différences. Les attitudes sont construites socialement pour et par la langue ; elles peuvent apparaître spécifiquement dans une langue et n'être pas ou mal identifiées par un apprenant de cette langue. Une expression peut ainsi leurrer l'apprenant qui a « reconnu » un faux-ami .

Nous commencerons par la présentation du corpus japonais sur lequel se base cette recherche. La première expérience perceptive, réalisée par des auditeurs japonais natifs, permet de valider les attitudes japonaises. Puis, nous présenterons (1) comment les apprenants français de niveau 0 en japonais perçoivent les attitudes japonaises (2) qu'est-ce que les apprenants français de niveau 1 en japonais ont appris sur les attitudes japonaises, et (3) quelles attitudes les apprenants français de niveau 2 en japonais savent ou non identifier.

2. SÉLECTION DES 12 ATTITUDES JAPONAISES

Nous avons retenu un ensemble de 12 attitudes représentatives du japonais selon la littérature de ce domaine (Erickson et al. [5], Ofuka et al. [10] etc.), mais surtout selon les méthodes d'enseignement du japonais. Il s'agit de : « *doute-incrédulité* », *« évidence »*, « *exclamation de surprise* », « *autorité* », « *irritation* », « *arrogance-impolitesse* », « *sincérité-politesse* », « *admiration* », « *kyoshuku* », « *politesse-simple* », « *déclaration* » et « *question-simple* ». Certaines de ces attitudes sont spécifiques ou particulières à la culture japonaise, notamment celles qui sont liées à la stratégie de politesse japonaise : « *politesse-simple* », « *sincérité-politesse* » et *« kyoshuku » vs.* « *arrogance-impolitesse* ». L'attitude de « *sincérité-politesse* » apparaît lorsqu'un locuteur considéré comme inférieur communique avec un interlocuteur considéré comme supérieur dans la société japonaise. Le locuteur inférieur exprime que son intention est sérieuse et sincère à travers cette attitude prosodique. L'attitude de « *kyoshuku* » (il n'existe pas d'entrée lexicale pour traduire ce terme en français) est une attitude typique de la culture japonaise : même si des situations d'interactions sociales semblables se produisent dans toutes les cultures, la langue japonaise a choisi de coder cette situation particulière en tant qu' « attitudinème ». Elle apparaît quand le locuteur est dans une situation où son statut social est inférieur à celui de son interlocuteur, et quand il a de plus un avis contraire ou qu'il désire lui demander un service. Elle est décrite par T. Sadanobu [11] comme "*a mixture of suffering ashamedness and embarrassment, (which) comes from the speaker's consciousness of the fact that his/her utterance of request imposes a burden to the hearer*" (Sadanobu [11] p.34).

3. CORPUS

Puisque notre objectif est de mesurer le comportement perceptif des auditeurs français, nous avons besoin des données de référence sur les attitudes japonaises. Les phrases utilisées doivent être dégagées d'informations lexico-syntaxiques sur les attitudes : seule l'information

attitudinale prosodique peut être utilisée, et les autres variations prosodiques doivent être équilibrées, afin de contrôler et de mesurer un éventuel biais dû à l'intonation japonaise ou à l'accent lexical qui pourraient être interprétés à tort par des auditeurs francophones comme des indices prosodiques de certaines attitudes.

La première étape consiste en l'enregistrement d'un corpus contrôlé, construit à partir des principes théoriques. La construction des paires minimales nous permet d'observer uniquement l'effet du facteur manipulé. Sur la base de tels corpus contrôlés, une analyse acoustique permettra ultérieurement de proposer un modèle morphologique des attitudes prosodiques du japonais. Notre corpus est basé sur sept phrases dont la longueur varie de 1 à 8 mores. La structure syntaxique des phrases est soit une phrase elliptique réduite à un mot simple, soit une structure simple de type « verbe-objet ». Pour les phrases de 8 mores, l'accent lexical se trouve sur la première, la deuxième ou la troisième more, ou la phrase est non accentuée. Afin d'exprimer certaines attitudes comme le doute ou la surprise, la voyelle [u] peut être insérée en fin de phrase, et dans ce cas l'accent lexical est réalisé sur la septième more. Les phrases ont été construites afin de n'avoir aucune connotation sémantique particulière dans quelle que soit la région du Japon. Chaque phrase est produite avec les douze expressions attitudinales. Un enseignant universitaire de japonais pour apprenants français, de langue maternelle japonaise, a été le locuteur de ce corpus. En effet, les apprenants d'une langue étrangère doivent acquérir les attitudes prosodiques de la langue et de la culture cibles, et les enseignants de langue savent produire différents types d'attitudes pour des raisons didactiques et pragmatiques. Le corpus contient en tout 84 stimuli, soit 7 phrases produites avec 12 attitudes différentes. Tous les stimuli sont employés pour le test perceptif.

Tableau 1 Corpus des attitudes japonaises : 7 phrases de longueur variée avec les différentes positions de l'accent lexical marquées par un astérisque.

Nb mora	Enoncé	Traduction
1	Me	L'oeil
2	**Na*** ra	Nara
5 (3+2)	**Na*** rade neru	Il dort à Nara
8 (4+4)	**Na*** goyade nomimas	Il boit à Nagoya
8 (4+4)	Na**ra*** shide nomimas	Il boit dans la ville de Nara
8 (4+4)	Matsu**ri*** de nomimas	Il boit à la soirée
8 (4+4)	Naniwade nomimas	Il boit à Naniwa

4. PROTOCOLE EXPERIMENTAL

La validation des attitudes japonaises est effectuée grâce à des auditeurs japonais natifs. 15 auditeurs japonais (11 femmes et 4 hommes) parlant le dialecte de Tokyo (est du Japon), dont l'âge moyen est de 29.5 ans, ont choisi une attitude parmi un choix fermé de 12. Le premier test de perception a été effectué auprès de 15 auditeurs français (10 femmes et 5 hommes) n'ayant jamais été confrontés à la langue japonaise. Ces 15 auditeurs sont classés dans le groupe de niveau 0, dont l'âge moyen est de 25.4 ans. Les auditeurs n'ont mentionné aucun trouble d'audition. Sur l'interface du test (en japonais ou en français), chaque attitude est définie et illustrée par un exemple de situation dans laquelle une telle attitude peut se produire. Aucun sujet n'a rapporté de difficulté de compréhension des définitions des attitudes.

Une deuxième expérience a été effectuée avec 16 auditeurs français, apprenants de japonais. Leurs compétences du japonais ont été évaluées comme homogènes, et du niveau 1 de notre grille de compétences (utilisée à l'Université Stendhal en particulier), qui est le niveau débutant : ils peuvent déjà parler et comprendre le japonais, mais avec des difficultés établies par le test de compétences. Nous avons présenté la même interface que celle utilisée pour les Français de niveau 0. Une troisième expérience a été menée avec 16 auditeurs français, tous évalués au niveau 2 : ils parlent couramment le japonais. Nous avons utilisé la même interface en français. Tous les sujets de ces expériences ont écouté chaque stimulus une fois seulement. Pour chaque stimulus, nous leur avons demandé de choisir l'attitude perçue parmi les 12 possibles. L'ordre de présentation des stimuli était aléatoire et différent pour chaque sujet.

5. RÉSULTATS

5.1. Validation avec les auditeurs japonais

Selon un test de khi-deux, les distributions des réponses de chaque attitude sont significativement différentes du hasard. Ensuite, nous avons examiné l'effet de la longueur sur le choix des attitudes. Ce test donne des distributions de réponses significativement différentes entre les phrases de deux et cinq mores, et celles de cinq et de huit mores. Il n'y a pas d'effet significatif de la position d'accent lexical sur le choix des attitudes. Afin de déterminer quelles attitudes ont effectivement été reconnues par les auditeurs, nous nous sommes basés sur le critère suivant : le taux d'identification moyen doit être au-dessus de deux fois le seuil du hasard. Selon ce critère, sept attitudes (i.e. *arrogance-impolitesse*, *déclaration*, *doute-incrédulité*, *politesse-simple*, *exclamation de surprise*, *irritation* et *question simple*) ont été reconnues sans confusion particulière. « *Autorité* » a été confondue avec « *évidence* », et « *évidence* » avec « *arrogance-impolitesse* ». L'évidence montre que le locuteur est sûr de lui-même et cette expression de la certitude peut parfois être perçue comme irrespectueuse vis-à-vis de l'interlocuteur. En effet, dans la société japonaise, même si on est sûr de soi, on évite d'afficher cette certitude car cela risque d'être interprété comme si le locuteur cherchait à s'imposer.

Les deux attitudes de politesse particulièrement japonaises (*sincérité-politesse* et *kyoshuku*) ont été confondues. Cette confusion s'explique mal acoustiquement, car leurs morphologies prosodiques sont différentes. Même s'il est vrai qu'une distance acoustique prosodique se mesure essentiellement par une distance perceptive, nous pensons qu'ici la confusion perceptive est liée à la proximité « sémantique » de ces deux valeurs socioculturelles, qui servent à montrer son humilité face à une personne socialement supérieure. Il faut noter aussi que « *sincérité-*

politesse » a été également confondue avec « *politesse-simple »*, ce qui n'est pas le cas de « *kyoshuku »*.

L'attitude d'*admiration* a été confondue avec « *politesse-simple »*. Ces deux attitudes sont interconnectées dans la société japonaise. Ce phénomène peut s'expliquer par la polysémie lexicale des items comme « *sonkee »* [admiration/politesse] et « *keifuku »* [admiration/politesse].

Tableau 2 Matrices de confusions en valeurs relatives pour 15 auditeurs japonais. Les valeurs dans les cellules diagonales qui sont reconnues au-dessus du seuil du hasard (plus de 16.6%), sont présentées en gras. Les attitudes qui présentent une confusion significative sont présentées en italique gras.

	Presented attitudes										
Percepted attitudes	*AD*	*PO*	*KYO*	*SIN*	*AR*	*AU*	*IR*	*DO*	*EX*	*QS*	*EV*	*DC*
Admiration	**21,9**	4,8	1,9	3,8	0,0	0,0	0,0	1,9	1,0	0,0	0,0	0,0
Politeness	***26,7***	**64,8**	2,9	***17,1***	1,0	4,8	0,0	0,0	0,0	8,6	1,0	11,4
Kyoshuku	11,4	9,5	**24,8**	***27,6***	0,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0
Sincerity	14,3	6,7	***26,7***	**32,4**	0,0	5,7	0,0	0,0	0,0	1,9	0,0	1,0
Arrogance	1,9	0,0	1,9	0,0	**72,4**	5,7	2,9	5,7	0,0	1,0	***21,0***	10,5
Authority	0,0	0,0	15,2	1,0	9,5	**51,4**	11,4	1,0	1,0	1,9	9,5	2,9
Irritation	1,0	0,0	12,4	1,0	5,7	4,8	**85,7**	13,3	4,8	1,0	2,9	1,0
Doubt	0,0	1,0	0,0	0,0	0,0	0,0	0,0	**56,2**	14,3	3,8	0,0	0,0
Surprise	9,5	1,9	1,0	1,9	0,0	0,0	0,0	14,3	**59,0**	1,0	4,8	0,0
Interrogation	0,0	1,0	0,0	1,9	0,0	0,0	0,0	7,6	14,3	**77,1**	2,9	0,0
Evidence	8,6	1,9	10,5	3,8	5,7	17,1	0,0	0,0	5,7	1,0	**45,7**	7,6
Declaration	4,8	8,6	2,9	9,5	5,7	10,5	0,0	0,0	0,0	1,9	12,4	**65,7**

5.2. Auditeurs français - niveau 0

La distribution des réponses pour chaque attitude est significativement différente du hasard. Un effet significatif de la longueur a été observé entre les phrases d'une et de deux mores. Il n'y a pas d'effet significatif de l'accent lexical sur le choix des attitudes pour les sujets français de niveau 0.

Selon le même critère d'identification appliquée précédemment pour les auditeurs japonais, les résultats suivants ont été obtenus. La figure 1. montre que : « *autorité »*, « *irritation »* et « *admiration »* ont été perçues sans confusion significative selon notre critère. Cependant, les sujets français de niveau 0 ont montré un taux d'identification faible pour l'attitude d'*arrogance-impolitesse*. Cette attitude a été confondue avec « *déclaration* » et « *autorité »*. Les auditeurs français de niveau 0 n'ont pas reconnu deux attitudes de politesse liées étroitement à la société japonaise (i.e. *sincérité-politesse* et *kyoshuku*). « *Sincérité-politesse »* a été reconnue comme « *politesse-simple »* ou « *kyoshuku »,* qui sont différents degrés de politesse. En revanche, l'attitude de « *kyoshuku »* a été reconnue comme « *irritation »*, « *arrogance-impolitesse »* ou « *autorité »*. Ce résultat était très attendu car ces attitudes n'apparaissent pas dans la société française, et surtout la qualité de voix de « *kyoshuku »* est utilisée en français uniquement dans des expressions d'émotions négatives.

Les auditeurs français ont confondu également « *question-simple »* avec « *déclaration »*. Ce résultat implique une possibilité pour les Français natifs de percevoir la prosodie d'une question comme celle d'une déclaration. Les sujets ont montré également une confusion réciproque significative entre « *déclaration »* et « *evidence »*, entre « *doute-incrédulité »* et « *exclamation de surprise »*, et encore entre « *politesse-simple »* et « *sincérité-politesse »*.

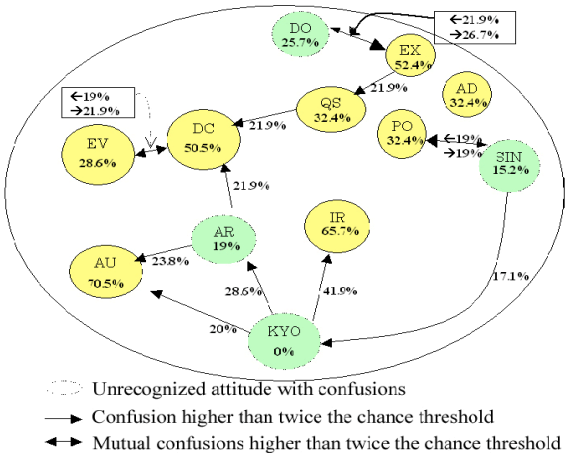


Figure 1 : Graphe de confusions des 15 auditeurs français niveau 0 : les pourcentages présentés en dehors des cercles indiquent le taux de confusion. D'autres pourcentages mentionnés en dessous des étiquettes de chaque attitude représentent les taux d'identification des attitudes. NOTE: AD(admiration), AR(arrogance-impolitesse), AU(autorité), DC(déclaration), DO(doute-incrédulité), EV(évidence), EX (exclamation de surprise), IR(irritation), KYO(kyoshuku), PO(politesse), QS(question-simple) and SIN(sincérité-politesse)

5.3. Auditeurs français - niveau 1

À ce niveau de japonais (débutants), les sujets ont appris à identifier « *sincérité-politesse »* et « *doute-incrédulité »*. Leurs confusions et erreurs d'interprétation ont changé. « *Sincérité-politesse »* a été confondue avec « *déclaration »*, et « *arrogance-impolitesse »* a également été confondue avec « *évidence »*. Il existe toujours une confusion mutuelle entre « *doute-incrédulité »* et « *exclamation de surprise »*. Ils ont appris à discriminer « *arrogance-impolitesse »* *et* « *autorité »*, « *politesse-simple »* *vs.* « *sincérité-politesse »*, « *déclaration »* *vs.* « *évidence »*. Cependant, ils ont encore confondu « *kyoshuku »* avec « *irritation »*, « *arrogance-impolitesse »* et « *autorité »*, il ont confondu également « *arrogance-impolitesse »* avec « *déclaration »*, et « *question-simple »* avec « *déclaration »*, ce qui peut être un handicap à la communication pour ces sujets qui commencent à parler le japonais.

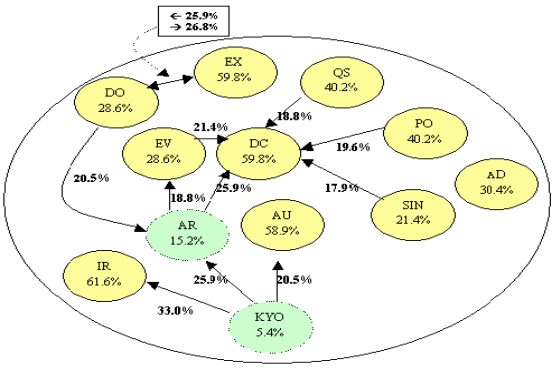


Figure 2 : Graphe de confusions pour 16 auditeurs français (niveau 1) : les pourcentages présentés en dehors des cercles indiquent le taux de confusion. D'autres pourcentages mentionnés en dessous des étiquettes de chaque attitude

représentent les taux d'identification d'attitude.

5.4. Auditeurs français – niveau 2

À ce niveau, où les sujets français sont à l'aise en japonais, deux attitudes qui étaient déjà confondues par les auditeurs du niveau inférieur restent mélangées. Il s'agit de « *arrogance-impolitesse* » et « *kyoshuku* ». « *Kyoshuku* » est toujours reconnu comme « *irritation* », « *arrogance-impolitesse* » ou « *autorité* », et « *arrogance-impolitesse* » comme « *déclaration* » ou « *évidence* ». Par contre, ces apprenants discriminent *question-simple* et *déclaration*. Les sujets confondent aussi « *déclaration* » avec « *sincérité-politesse* » et aussi *« doute-incrédulité »* et *« exclamation de surprise »* (ce qui n'est pas le cas pour les auditeurs japonais et pour les Français avec les attitudes françaises de « *doute* » et « *surprise* » [1]).

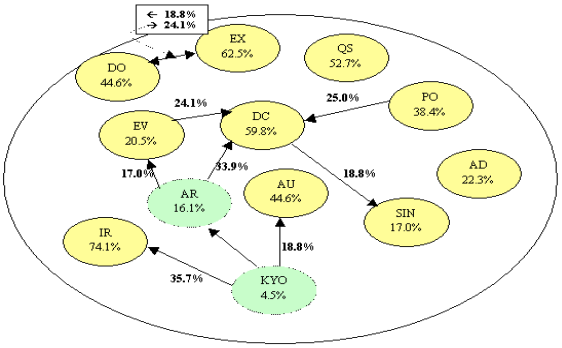


Figure 3 : Graphe de confusions pour 16 auditeurs français (niveau 2): les pourcentages présentés en dehors des cercles indiquent le taux de confusion. D'autres pourcentages mentionnés en dessous des étiquettes de chaque attitude représentent les taux d'identification des attitudes.

CONCLUSION

D'une manière générale certaines attitudes sont systématiquement reconnues, même par les auditeurs français naïfs en japonais. Aucun effet de l'accent lexical n'a été observé, ni pour les auditeurs japonais, ni pour les auditeurs français, quel que soit leur niveau de langue. Nous avons analysé les paramètres acoustiques de la déclaration et de la question-simple afin de chercher une explication à leur confusion par les auditeurs français, qui ne disparaît qu'au niveau 2. Mais nous n'avons pu observer aucune similitude, selon les critères de la prosodie du français, entre « déclaration » et « *question-simple* » portant différents accents lexicaux en japonais, ni entre les prototypes japonais et français de question/déclaration. Ceci fera l'objet d'une étude spécifique ultérieure. Certaines confusions rencontrées à des niveaux d'apprentissage différents par les Français sont typiquement des cas de « faux-amis » prosodiques qu'il est nécessaire d'identifier lors d'une tâche d'apprentissage du japonais par des Français. Parallèlement à ces résultats, une analyse acoustique du corpus devrait indiquer les caractéristiques prosodiques de chaque attitude. Il est important d'examiner les confusions dues aux différences interculturelles avec des stimuli composés de phrases françaises et d'attitudes prosodiques japonaises superposées. D'autres expériences sont en cours : un test de perception sur le paradigme du « gating » afin d'établir quand arrive la prédiction des attitudes [1] ; une observation des comportements perceptifs de sujets américains natifs de niveau 0 à 3 en japonais, parallèlement à l'étude croisée de sujets japonais sur le français et l'anglais.

REMERCIEMENTS

Nous remercions particulièrement T. Sadanobu, de l'université de Kobé et Nick Campbell, ATR, Japon.

BIBLIOGRAPHIE

[1] Aubergé, V., Grépillat, T., Rilliard, A.: Can we perceive attitudes before the end of sentences ? 5th Eurospeech, (1997) 871-874.

[2] Ayuzawa, T.: Nihongo no gimonbun no inritsuteki tokucyou. In Nihongo no inritsu ni mirareru bogo no kansyou (2), Grand-in-aid for Sci. Re. on Priority areas (D1), research rep. 1992 (1992) 1-20.

[3] Campbell, N.: Modelling Affect in Speech Communication, Beiing (2003).

[4] Erickson, D., Ohashi, S., Makita, S., Kajimoto, N., Mokhtari, P.: Perception of naturally-spoken expressive speech by American English and Japanese listeners. In CREST International Workshop on Expressive Speech Processing (2003) 31-36.

[5] Ito, M.: The Contribution of Voice Quality to Politeness in Japanese. In VOQUAL'03, Geneva (2003) 157-162.

[6] Ko, M.: Teinei hyougen ni mirareru nihongo onsei no inritsuteki tokucyou. In the Phonetic Society of Japan 1993 Annual Convention (1993) 35-40.

[7] Matsumoto, E., Sadanobu, T.: Nihongo no inritsu niokeru rikimi to, nihongo gakushuusha no rikaido. In: Departement of Japanese Studies, The Chinese University of Hong Kong and Society of Japanese Language Education (eds.): Quality Japanese Studies and Japanese Language Education in Kanji-Using Areas in the New Century, Hong Kong Himawari Publishing Co. (2001) 455-461.

[8] Morlec, Y., G. Bailly, and V. Aubergé Generating prosodic attitudes in French: data, model and evaluation. Speech Communication, (2001) 33(4), 357-371.

[9] Ofuka, E., McKeown, J.D., Waterman, M.G. , Roach, P.J.: Prosodic cue for rated politeness in Japanese speech. In Speech Communication, 32 (2000) 199-217.

[10] Sadanobu, T.: A natural history of Japanese pressed voice. In Journal of the Phonetic Society of Japan, Vol.8. No.1 (2004) 29-44.

[11] Van Bezooijen, R.: Sociocultural aspects of pitch differences between Japanese and Dutch women. In Language and Speech 38, 3 (1995) 253-266.

Expressions hors des tours de parole : éthogrammes du « *feeling of thinking* »

Fanny Loyau, Véronique Aubergé & Anne Vanpé

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
{loyau, auberge}@icp.inpg.fr

ABSTRACT

During our collect of an expressive corpus, a large quantity of non verbal information has been registered too: top body and face movements, and voice events. We are particularly interested by only these actions which happen outside the talk turn, when the subject thinks, and feels about what he thinks. We want to know if these events are real indices of signals about the mental states or the affective states of the subjects. For that, a typical ethogram methodology has been applied to label these non speech parts into primitive icons of top body movements, face movements and voice events, in order not to take any decision about the interpretation of what could be expressed by these events, but to classify variant movements into minimal icons.

1. INTRODUCTION

Dans le domaine de l'interaction verbale et de la communication expressive, des études de plus en plus nombreuses sont consacrées non seulement aux expressions transmises par le locuteur pendant son tour de parole, mais également aux informations émises par le sujet humain en dehors de son tour de parole, en particulier quand le sujet suit en ligne l'interlocuteur qui est en train de dérouler son tour de parole (« feedback », [3][5]). Il peut alors lui renvoyer des informations sur son attention, l'état de son traitement mental – sa compréhension – sur ce qu'il reçoit de l'interlocuteur, ses opinions sur ce qu'il reçoit, les émotions que ces traitements induisent sur lui. En dehors de son tour de parole, le sujet peut également être dans une situation de traitement d'une tâche cognitive et / ou physique à accomplir, et faire ou laisser apparaître des informations sur ses états mentaux et affectifs dans le traitement de cette tâche. Cette situation apparaît en particulier fréquemment dans les interactions personne-machine.

Cet article présente les analyses préliminaires des expressions multimodales d'un corpus expressif multi-locuteurs (Sound Teacher de E-Wiz [1]) dans les parties de l'interaction où les sujets ne sont pas dans leur tour de parole, et dans lesquelles cependant les expressions dans la voix, la face ou le corps sont nombreuses et variées. Nous proposons ici les grandes lignes d'une méthode d'annotation de ces expressions qui n'est pas basée sur une mesure automatique de l'image ou du signal vocal, mais qui restreint le rôle de l'expert humain à la détection d'icônes gestuelles ou vocales minimales, sans interprétation de contenu informatif, rejoignant en cela une démarche classique d'éthogramme.

2. LE CORPUS EXPRESSIF SOUNDTEACHER /E-WIZ

Le corpus expressif Sound Teacher de E-Wiz [1] a été réalisé à partir de l'enregistrement de 17 sujets, 11 femmes et 6 hommes, placés dans une situation d'apprentissage des voyelles des langues du monde à l'aide d'un pseudo système révolutionnaire, Sound Teacher. Ces sujets sont « piégés » par un scénario de type magicien d'Oz : le sujet pensait communiquer avec un ordinateur, alors qu'en fait le comportement apparent de l'application est géré à distance par le magicien.

Le scénario se déroule en trois grandes phases, la première, dite d'entraînement, familiarise et rassure le sujet, une deuxième phase implique le sujet dans des tâches très simples sur lesquelles il est félicité, ce qui a induit chez l'ensemble des sujets des émotions globalement positives, et une troisième phase, de plus en plus complexe, dans laquelle sont renvoyés au sujet des jugements négatifs, qui se termine par une répétition de la tâche initiale simple, mais en retournant aux sujets de (faux) résultats très mauvais qui les ont soit fortement inquiété, soit déstabilisé. Après chaque enregistrement, les sujets ont auto-annoté leur production en notant selon leur propre choix (langage, dessins, signes etc) leurs états mentaux et affectifs finement au fil de l'avancement du scénario. Les sujets interagissent seulement par la parole, pour les réponses ou pour les phases de commentaires libres (pas de clavier ni souris). Ils sont isolés en chambre sourde face à un écran, et ne se savent pas enregistrés. La machine dialogue soit par du texte, soit par l'exécution de la demande de tâche. Les sujets sont donc alternativement en phase de lecture, réflexion, production de parole par proposition verbalisée de réalisation de la tâche (sous forme d'un mono-mot mono-syllabique).

3. ETIQUETAGE DU CORPUS : UN ÉTHOGRAMME

3.1. Du "feeling of knowing" au "feeling of thinking"

Sound Teacher est une situation de dialogue minimale, puisque le sujet sait que ses tours de parole ne changent pas la nature de l'interaction. La phase de communication humaine ou humanisée dans laquelle le sujet auditeur envoie un feedback à son interlocuteur « intentionnellement » n'est donc pas attendue. Pourtant, nous le montrons plus loin, pendant le « tour de parole » de la machine (lecture), le sujet affiche des expressions

riches pendant la dynamique de la lecture. Surtout, pendant la phase de préparation de leur proposition verbale, les sujets expriment à la fois des affects et des états mentaux. Dans une tâche encore plus spécifique que celle de Sound Teacher, (un sujet se voit poser une question de culture générale et n'arrive pas à fournir la bonne réponse ; le sujet sait pourtant qu'il connaît cette réponse, il l'a stockée dans sa mémoire, et pourra la retrouver plus tard, mais dans l'instant présent elle n'est pas disponible) des expressions révélant le processus mnésique du sujet ont été observées, étudiées et regroupées comme *feeling of knowing* [6]. La tâche Sound Teacher révélant des processus cognitifs et affectifs plus larges que la tâche mnésique exprimée en *feeling of knowing*, ils seront regroupés ici dans une phénoménologie plus générique que nous appelerons « *feeling of thinking* », expressions des états affectifs et mentaux.

3.2. Méthodologie

Le problème crucial posé dans cette étude est celui de l'annotation des expressions. Le scénario étant connu, les « états mentaux et affectifs » étant étiquetés par les sujets (et vérifiés pour certains dans des expériences perceptives), la subjectivité de l'étiquetage par un « expert » humain est d'autant plus grande. Nous avons donc fait le choix que les experts (deux experts pour 17 sujets) n'aient pas connaissance a priori des annotations des sujets. Le but est d'utiliser leurs compétences d'humains communicants pour dégager une icônicité minimale des signaux, mais en minimisant leur compétence interprétative (ils ne doivent pas être un participant humain ajouté à l'interaction, mais conserver une distance « objectivante »). Ils doivent se ramener le plus possible à un étiquetage de la « syntaxe icônique » des mouvements et événements vocaux, sans interprétation « sémantique » de l'expression (par exemple, pas d'étiquetage des gestes faciaux en sourire ou autres moues, mais icônes de géométrie et dynamique jugées différentes). Cette démarche, qui s'ancre dans une méthodologie d'éthogramme, est donc fondamentalement déterminée par les icônes minimales définies comme étant les étiquettes à poser sur le corpus. Une difficulté supplémentaire est introduite par la non généricité de certaines icônes que nous avons été amenés à décrire, sans que nous puissions a priori décider s'il s'agit d'une variante d'une icône générique (i.e. partagée par tous les sujets, susceptible d'être un signal communicatif) ou idiosynchrasique (i.e. spécifique à un sujet mais néanmoins indice « récupéré » de communication).

Une démarche éthologique

Pour ce faire, nous avons appliquer un protocole issu de l'éthologie (étude des mœurs et du comportement individuel et social des animaux domestiques et sauvages) : nous avons choisi d'annoter nos corpus à l'aide d'éthogrammes. Un tel objet représente l'inventaire des comportements d'une espèce. Plus précisément, « l'éthogramme consiste en un répertoire d'actes et de postures observés et définis de façon précise par l'expérimentateur ; la grille d'observation est construite d'après cet éthogramme et permet de quantifier la fréquence des comportements sur une période de temps donnée avec, éventuellement, leurs durées et enchaînements. Chaque intitulé est défini selon des critères de direction, de sens, de localisation, de distance, d'intensité ou d'amplitude». Ainsi, nous pouvons étiqueter nos parties des corpus sans parole en utilisant des icônes primitives pour les mouvements du haut du corps, ceux de la face, et les événements vocaux, sans avoir à prendre de décision qui serait du niveau de l'interprétation. Par exemple, le mouvement de la bouche présenté dans la figure 1 ne sera pas étiqueté en tant que « sourire » mais juste en tant qu'icône : « monter le coin des lèvres », que l'on appelle IGS, avec comme variables l'intensité, la durée et l'ouverture ou non de la bouche.

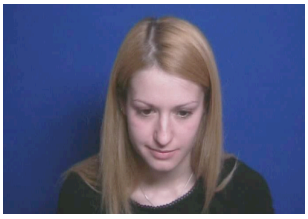


Figure 1 : IGS « monte le coin des lèvres - dissymétrie droite / faible / rapide / fermée ».

Dans la figure 2, se trouve un extrait d'une durée d'environ 25 secondes de l'étiquetage d'un des corpus, avec la partie de l'éthogramme correspondant, où se trouvent les descriptions de chaque icône utilisée dans ce bout d'étiquetage. Chaque occurrence d'icône est numérotée, pour pouvoir ensuite faire des analyses quant à la fréquence d'apparition de chaque icône.

- « eu » (5mn15) : dit « jaune » à 5mn17, IGG11 rapide
- « u » (5mn22) : dit « jaune » à 5mn23 IGF5 tenu
- « é » (5mn25) : dit « vert » à 5mn26 IGF6 tenu
- « o » (5mn30) : IGT3 droite / assez net / rapide, dit « jaune » à 5mn31 puis IGRe1 dans la diagonale, en bas à droite / rapide avec IGSo2 rapide
- « e » (5mn35) : IGLP5 à droite sur ce « e », dit « jaune » à 5mn36 avec IGS3, bouche quasi fermée / long puis IGSo4 rapide pendant lecture consignes

IGF : front plissé
IGLP : lèvres pincées
IGRe : regard hors de l'écran
IGS : coins de la bouche relevés
IGSo : sourcils haussés
IGT : tête penchée sur le coté

IGG : regroupement de plusieurs IG

Figure 2 : Partie de l'étiquetage et son éthogramme associé.

3.3. Signaux vs. indices

Certaines icônes vont sembler se retrouver chez tous les sujets, ce serait donc des signaux, tandis que d'autres icônes semblent propres à un sujet, on parlerait alors d'indices. Mais aucune décision n'est prise a priori sur les différents événements (gestes, face, voix), qui seront plus tard identifiés comme étant soit des signaux de communication, soit des indices biologiques, idiosyncrasiques dont la variabilité est associable à des changements d'états affectifs [2].

4. PREMIERS RÉSULTATS

Les micro et macro organisations temporelles sont fondamentales, soit quand elles sont cohérentes avec l'évolution des états affectifs des sujets, soit parce qu'elles sont révélatrices de ces états (certaines icônes ne seront pas directement porteuses d'information, mais l'organisation temporelle de ces icônes le sera [2]).

4.1. Organisation temporelle

En moyenne, chaque sujet a donné lieu à un corpus d'environ 40 minutes (figure 3).

Le temps alloué aux moments de parole est lui d'approximativement 8 minutes, il y a donc 80% du temps de communication qui se situe hors des tours de parole.

Durant les 32 minutes utilisées par le sujet pour lire les consignes (8 minutes) et surtout pour penser aux réponses qu'il devra ensuite oraliser (24 minutes), nous n'avons pas trouvé de position « neutre » pour le haut du corps, le visage, ni de moment totalement silencieux.

Il se passe toujours quelque chose, et c'est ce que nous avons essayé d'étiqueter, de la façon la plus objective possible.

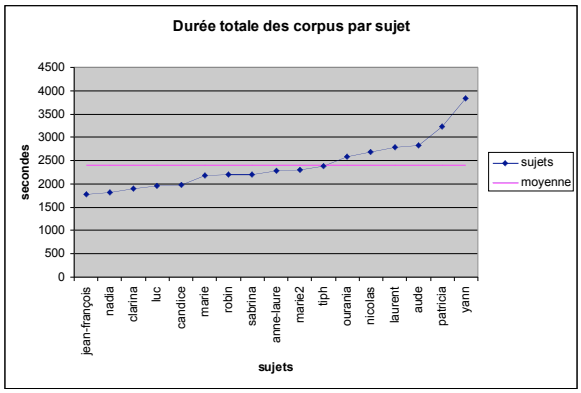


Figure 3 : Durée moyenne des 17 corpus.

Les caractéristiques générales comme les temps de réponse ont été traitées pour tous les sujets, mais pour l'instant l'analyse détaillée présentée ensuite n'a été menée que pour 5 d'entres eux.

Temps de réponse

Le temps moyen de réponse, correspondant au temps entre le moment où le sujet entend les stimuli et celui où il parle pour donner sa réponse est de 4,5 secondes. Cette durée est plus importante pour la phase d'entraînement, diminue dans la phase suivante, positive, puis se rallonge à nouveau dans la dernière phase, regroupant induction négative et déstabilisation (figure 4).

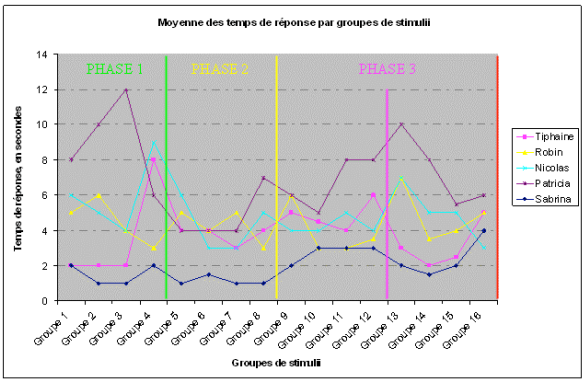


Figure 4 : Moyenne des temps de réponse, par phases.

L'écart type est nettement plus important pour les phases d'entraînement et négative que pour la phase positive, assez stable.

4.2. Expressions du « backchannel »

Nous nous intéressons tout particulièrement au lien qu'il va pouvoir y avoir entre les occurrences des différents indices apparaissant hors des tours de parole et à la fois l'auto annotation faite par les sujets eux-mêmes et les phases du scénario.

Commun aux sujets d'E-Wiz

Voici des mouvements que l'on retrouve chez tous les sujets d'E-Wiz : IGF « plisser le front », IGSo « hausser les sourcils », IGSof « froncer les sourcils », IGY« plisser les yeux », IGRe « regards hors de l'écran », IGN « plisser le nez », IGLp « passer ses lèvres l'une sur l'autre », IGLm « mordre sa lèvre », IGM « plisser le menton ».

Situation négative

Les regards hors de l'écran : Les sujets regardent tous parfois hors de l'écran, dans toutes les phases, mais surtout lorsqu'ils préparent leur réponse, et plus souvent là encore dans la phase négative. Les moments où ont lieu ces regards sont étiquetés par les sujets eux-mêmes avec en majorité les termes suivants : « perplexité », « doute », « stress », « ennui », « approximation », « perplexité », « incompréhension », « agacement »… Lorsqu'on s'intéresse aux regards qui ont lieu pendant les tours de parole, cela corrobore ces résultats : il y en a plus lors des dernières phases, on retrouve les mêmes étiquettes négatives données par les sujets.

Les rires : contrairement à ce qu'on aurait pu croire a priori, les rires apparaissent plus en situation négative qu'en situation positive. Ils correspondent alors à différentes étiquettes comme « fatigué et amusé », « irrité, anxieux », « stressé, je ne comprends pas » « surpris, nerveux » « doute, très agacé, ri de ma mauvaise performance », « rire = tentative de décontraction », « déception mais m'en amuse », « au pif, une envie de rigoler ». Le rire semble ici avoir pour fonction de permettre au locuteur de changer d'état, de ne pas rester dans une situation négative désagréable. La majorité des

rires ont toutefois lieu pendant les tours de parole, mais aussi lors de situations négatives.

La protrusion des lèvres : cet indice, s'il est présent dans toutes les phases, l'est, lui aussi, plus dans la troisième. Il correspond également à des auto annotations négatives : « j'ai l'air déçue », « concentré », « agacé », « concentration, ennui ». Il n'y en a pas du tout lors des tours de parole.

Les bruits de bouche : Ces bruits, comme des sifflements, des fricatives, des plosions, sont plus nombreux et plus irréguliers dans la phase négative, et augmentent tout particulièrement dans la partie de déstabilisation de cette dernière en devenant également de plus en plus irréguliers. Très peu ont lieu lors des tours de parole.

Ces comportements sont cohérents avec les résultats plus généraux concernant les indices biologiques du comportement observés chez les joueurs de tennis lors de situations inconfortables, de désarroi [2].

Situation positive

Il semble, pour l'instant, que très peu d'indices soient propres à cette situation. Pourtant, les deux premières phases, suivant le scénario, sont censées être positives. Mais dans l'auto annotation faite par les sujets, peu d'étiquettes sont positives, même dans ces deux phases. Les termes les plus fréquents sont les suivants : « ennui », « agacement », « doute ».

Le terme de « concentration » est lui aussi utilisé souvent, par tous les sujets, mais selon les autres mots auxquels il est relié il aura une connotation positive (dans les premières phases), ou négative (dans les dernières) : « plus de sérieux, de concentration », « grande concentration le but étant de comprendre ce qui est prononcé » / « ennui, concentration », « concentration – agacement ».

Les sourires : ils sont plus fréquents dans les deux premières phases, et concordent aux étiquettes suivantes : « calme », « assez calme », « concentrée » « fier, content », « étonné et doute », « très fier, content, étonné ». Ils sont donc utilisés à des moments très différents de ceux des rires. Pendant les tours de parole, cela diffère fortement d'un sujet à un autre : chez un il y en a beaucoup pendant les premières phases, chez un autre il y en a plus qu'hors des tours des tours de parole, et ce pour chaque phase, chez deux autres sujets il n'y en a quasiment pas, 2 dans la deuxième seulement pour l'un de ces sujets, enfin chez le cinquième sujet les sourires pendant le tour de parole ne sont pas majoritaires mais ont lieu principalement lors des dernières phases.

Spécifique au sujet

En ce qui concerne les événements observé chez un seul sujet, il y a l'icône « penche la tête de coté », propre donc à un seul sujet, qui fait ce mouvement principalement avant de répondre à un stimulus sauf dans la sous partie de déstabilisation, et très majoritairement sur sa droite (figure 5).

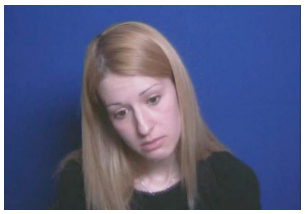


Figure 5 : icône IGT « tête penchée côté droit ».

Cette icône, qui apparaît dans toutes les phases, figure dans des situations positives, étiquetées par le sujet comme « calme », « assez calme », « concentration », mais surtout dans des moments négatifs, étiquetés comme « mal à l'aise », « oppressée », « inquiétude »…Un autre sujet est le seul à renifler, surtout avant de donner sa réponse, et uniquement dans la phase d'entraînement. Dans l'auto annotation cela correspond à des parties où ce sujet dit être « concentrée », ou « concentrée, stressée ».

5. CONCLUSION

Dans ce travail très préliminaire, nous avons essayé d'ébaucher une méthodologie encore empirique d'éthogrammes des comportements expressifs des sujets dans l'interaction mais hors de leur tour de parole. Nous avons déjà pu observer que l'organisation temporelle joue un rôle fondamental, aussi bien au niveau global que local. Nous allons confronter les icônes que nous avons identifiées d'abord à une validation statistique de leur pertinence, puis à leur validation perceptive, à la fois dans des tests de perception d'icônes isolées, ou de schémas rythmiques isolés, et en synthèse avec l'agent conversationnel, « GRETA » [4].

BIBLIOGRAPHIE

[1] V. Aubergé, N. Audibert, and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, 179-182, 2004.

[2] G, Carlier, and C. Graff, to be published. Unpredictability as a counter strategy: An analysis of elite matches. *Journal of Sciences*, 2006.

[3] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. *IVA'05 International Working Conference on Intelligent Virtual Agents*, 2005.

[4] I. Poggi, C. Pelachaud, F. de Rosis, V. Caroglio, B. de Carolis. GRETA. A Believable Embodied Conversational Agent. *Multimodal Intelligent Information Presentation*, O. Stock and M. Zancarano, eds, , Kluwer, to appear, 2005.

[5] M. Schröder, D. Heylen and I. Poggi, to be published. Perception of non-verbal emotional listener feedback. *Speech Prosody 2006*.

[6] M. Swerts and E. Khramer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 2004.

Acquisition de la liaison chez l'enfant francophone : formes lexicales des Mots 2

Dugua Céline & Chabanal Damien

Lidilem (Université Stendhal, Grenoble 3) ; LRL (Université Blaise Pascal, Clermont 2)
celine.dugua@u-grenoble3.fr ; damien.chabanal@univ-bpclermont.fr

ABSTRACT

The aim of this research is to establish which lexical forms in Word2 position are available to the child in liaison processes. It is known that the child makes many errors at the beginning of the acquisition of liaison and elision, of the sort "la noreille, le zours, petit néléphant, un zéléphant…". These variants argue in favour of the encoding of the liaison consonant at the beginning of Word2 in the early stages of the acquisition of liaison. To test this hypothesis, we present the results of two studies (one transversal, the other longitudinal) of French-speaking children between 2;6 and 6;3. The data collected allow us to conclude that the children have, at the outset, in their mental lexicon, Word2 forms with both an initial consonant and an initial vowel. These observations suggest that the child posseses several exemplars of the same Word2.

1. INTRODUCTION

La liaison est un phénomène d'alternance phonologique qui s'actualise entre deux mots (Mot 1 et Mot 2) par l'apparition d'une consonne de liaison dont la nature (/n/, /z/, /t/ dans 99.7% des cas, Boë, & Tubach [1]) est déterminée par le type de Mot 1. Les enfants sont confrontés à une double difficulté relative à la liaison. D'une part, comme dans l'enchainement, la frontière lexicale et la frontière syllabique ne concordent pas (*deux ours* segmenté [dø.zurs]). L'enfant en contact avec sa langue, sans la référence de l'écrit, hésiterait logiquement entre trois types de stratégies de segmentation : la segmentation [dø] / [zurs] (rattachement de la consonne au Mot 2), la segmentation [døz] / [urs] (rattachement de la consonne au Mot 1) ou la non segmentation [døzurs]. Nous souscrivons à l'hypothèse selon laquelle le jeune locuteur confronté à cette ambiguïté privilégiera les séquences CV et choisira la première segmentation. Selon MacNeilage & Davis [2] des caractéristiques bio-mécaniques engendreraient cette préférence pour CV. La production des formes CV serait contrainte par le cycle mandibulaire chez l'enfant. En outre, Vallée, Rousset & Boë [3] observent que les syllabes CV sont les plus fréquentes dans un grand nombre de langues du monde. L'enfant encodera donc la consonne de liaison à l'initiale de la représentation lexicale du Mot 2, c'est-à-dire celui qui suit la position de la consonne de liaison. Toutefois, cette stratégie de segmentation n'est pas incompatible avec une phase peut-être plus précoce lors de laquelle l'enfant ne segmenterait pas les séquences avec liaison et récupèrerait donc dans son lexique des constructions plus complexes (Bybee [4]). D'autre part, la seconde difficulté relative à la liaison est liée à la variation de la consonne de liaison en fonction de la nature du Mot 1. En mettant cette caractéristique en parallèle avec la prégnance de la segmentation CV, Chevrot & Fayol [5], Chabanal [6] et Dugua [7] suggèrent que l'enfant dispose de plusieurs allomorphes du même terme qu'il utilise variablement, tels que /zurs/, /turs/, /nurs/ pour le Mot 2 *ours*. Chabanal [6] remarque que les enfants âgés entre 2;6 et 3;4 produisent au cours d'une même séance d'enregistrement les Mots 2 sous les formes [lurs], [nurs], [zurs], [turs], [urs] à la suite de l'adjectif "petit". L'enfant pourrait disposer parfois de cinq allomorphes dans son lexique interne mais ne parviendrait pas immédiatement à faire le lien avec le Mot 1, ce qui expliquerait des activations hasardeuses et les variations que l'on a pu noter.

2. TACHE D'APOSTROPHE

Afin d'observer la disponibilité des variantes de Mots 2 chez les jeunes enfants en phase d'acquisition, nous avons mis en place un protocole expérimental sur la base d'une tâche d'apostrophe.

Cette tâche s'inscrit dans deux approches méthodologiques complémentaires : une étude transversale et une étude longitudinale. Le principe général de la tâche d'apostrophe est de faire produire à l'enfant des noms en isolation, sans déterminant. Ainsi, nous nous attendons à recueillir des formes à consonne initiale telles /nurs/, /zurs/ issues de la segmentation de séquences CV.

2.1. Protocole

Le protocole est identique pour les deux études. Nous avons utilisé sept figurines d'animaux : quatre dont les noms sont à voyelle initiale (*âne*, *écureuil*, *éléphant, ours*) et trois dont les noms sont à consonne initiale (*cochon*, *chien, perroquet)*, ces derniers jouant le rôle de distracteur. Dans un premier temps, l'enfant devait choisir une figurine parmi les trois distracteurs, figurine qui servirait d'intermédiaire aux paroles de l'enfant.

Dans un deuxième temps, l'expérimentateur installait les figurines restantes côte à côte, face à celle choisie par l'enfant. La consigne consistait, pour l'enfant, à appeler ses amis les animaux par leur nom ; l'expérimentateur illustrait la consigne avec l'apostrophe d'une figurine distracteur : *chien, viens ici !* A chaque apostrophe, l'expérimentateur faisait avancer la figurine appelée vers l'enfant. Chaque figurine était appelée deux fois, par conséquent, nous disposons de deux productions pour chaque mot, soit huit productions par enfant. En outre, l'enfant choisissant l'ordre d'apostrophe des figurines, l'ordre de production est donc propre au choix de chaque enfant.

2.2. Etude transversale

Sujets

Les sujets de l'étude transversale sont 178 enfants âgés de 2;6 à 6;1 (Table 1) pris dans un échantillon de 200 enfants. Nous avons réduit notre échantillon initial afin d'aligner les tranches d'âge de ces enfants avec les âges des enfants de l'étude longitudinale (*infra*). Il s'agit d'enfants "tout venant" choisis dans différentes écoles de la région Rhône-Alpes. La durée de la passation variait, selon l'âge des enfants, entre 5 et 10 minutes environ.

Table 1 : Répartition des sujets de l'étude transversale en cinq tranches d'âge.

	Echelle d'âges	Moyenne	Effectif
Age 1	2;6-3;1	2;10	50
Age 2	3;6-4;3	3;10	33
Age 3	4;4-4;9	4;6	32
Age 4	4;10-5;5	5;1	28
Age 5	5;6-6;1	5;9	35

Résultats

Tout d'abord, nous avons recueilli des productions avec déterminant qui ne correspondent donc pas à des productions isolées telles que nous les attendions. Il s'agit de deux types de productions : "un + Mot 2" (eg : *un ours !*) et "l' + Mot 2" (eg : *l'âne !*). Nous considérons les séquences à initiale en /l/ comme des séquences précédées de déterminant car nous n'avons aucun moyen de savoir si le /l/ est un déterminant ou le phonème initial de la séquence /lurs/. Les productions avec déterminant sont présentes dans les quatre tranches d'âge et représentent globalement 24% des productions (Table 2).

Les variantes à initiale vocalique (variantes #V, eg : *ours, viens ici !*) représentent, toutes tranches d'âge confondues, 60% de l'ensemble des productions. Elles sont de plus en plus souvent produites par les enfants au fil des tranches d'âge. La courbe développementale augmente significativement entre les deux premières tranches d'âge (Test U de Mann-Whitney : U = 998, p = 0.0068), puis elle poursuit son augmentation régulièrement et non significativement jusqu'à la tranche d'âge 5 où elle atteint plus de 80% des productions.

Dans les variantes à consonne initiale (variantes #C), les consonnes impliquées sont majoritairement la consonne /n/, et dans une moindre mesure les consonnes /z/ et /t/ (respectivement seules 10 et 7 productions de ce type ont été relevées). Autrement dit, pour appeler la figurine de l'ours, l'enfant dit par exemple *nours, viens ici !* Chez les plus jeunes (2;6-3;1), il est à noter que 39.2% des productions sont à consonne initiale (Table 2). Ce type de production n'est donc pas marginal entre 2;6 et 3;1. La diminution de ces productions est significative entre les tranches d'âge 1 et 2 (Test U de Mann-Whitney : U = 1022.5, p = 0.0027), puis elle se stabilise pour atteindre un taux de 3.8% dans la dernière tranche d'âge.

Table 2 : Résultats de l'étude transversale (moyennes et écarts-types).

#V : forme à initiale vocalique (e.g. *ours !*) ; #C : forme à initiale consonantique (e.g. *nours !*) ; Avec déterminant (eg. *un ours ! l'ours !*).

	Variantes #V	Variantes #C	Avec déterminant
2;6-3;1	32.0% (*30.2*)	39.2% (*30.2*)	28.8% (*32.3*)
3;6-4;3	56.8% (*40.6*)	18.6% (*25.9*)	24.6% *(34.4)*
4;4-4;9	68.0% (*36.9*)	13.3% (*19.3*)	18.8% *(29.8)*
4;10-5;5	73.3% (*34.2*)	12.5% (*22.6*)	14.2% *(28.5)*
5;6-6;1	81.2% (*27.1*)	3.8% (*11.1*)	15.0% (*24.2*)

Discussion

Les résultats de la tranche d'âge 1 doivent attirer notre attention. En effet, la répartition entre les trois types de productions est homogène (selon le test non paramétrique de Wilcoxon, pas de différence significative entre ces trois types de productions dans cette tranche d'âge). Cette observation suggère que les jeunes enfants disposent, dans leur lexique, d'une variété de formes de mots et/ou séquences plus complexes. En d'autres termes, et en s'attachant aux Mots 2 de notre étude, ceux pouvant intégrer des contextes de liaison, il apparaît que les enfants possèdent à la fois des formes de mots à voyelle initiale, des formes de mots à consonne initiale et des séquences avec déterminant non segmentées.

2.3. Etude longitudinale

Sujets

Vingt enfants ont été suivis pendant 4 ans, à raison de cinq observations, entre les âges moyens de 2;10 et 5;9.

Table 3 : Echelle d'âges et moyennes des sujets de l'étude longitudinale lors des cinq observations.

	Echelle d'âges	Moyenne
Observation 1	2;6-3;1	2;10
Observation 2	3;6-4;3	3;11
Observation 3	4;4-4;9	4;6
Observation 4	4;10-5;5	5;2
Observation 5	5;6-6;3	5;9

Résultats

Les productions avec déterminant sont présentes dans les cinq temps d'observation. Lors du temps d'observation 4, le taux de productions avec déterminant est surprenant et semble faire exception au profil général de ce type de production.

Lors du premier temps d'observation, le taux de variantes à voyelle initiale représente la moitié des réalisations des enfants. Ce taux augmente ensuite avec un pic entre les temps d'observation 3 et 4 (Wilcoxon : z=-3.014, p=0.0026). Lors du dernier temps d'observation (entre 5;6 et 6;3), le taux de variantes à voyelle initiale atteint 81.8%.

Les variantes à consonne initiale impliquent majoritairement la consonne /n/ (102 productions en /n/ sur 130 productions à consonne initiale). Elles sont stables dans les deux premiers temps d'observation, puis elles diminuent de façon significative entre les temps 2 et 3 et entre les temps 4 et 5 (respectivement, Wilcoxon : z=-2.073, p=0.0382 ; z=-2.214, p=0.0269). Entre 5;6 et 6;3 plus aucune variante à initiale consonantique n'est produite.

Table 4 : Résultats de l'étude longitudinale (moyennes et écarts-types).

	Variantes #V	Variantes #C	Avec déterminant
2;6-3;1	50.0% *(33.5)*	28.6% *(30.9)*	28.2% (*38.1*)
3;6-4;3	47.8% (*36.3*)	28.6% (*32.1*)	23.6% *(28.7)*
4;4-4;9	58.3% (*37.4*)	19.6% (*28.5*)	22.1% *(26.7)*
4;10-5;5	84.4% (*20.6*)	7.8% (*16.3*)	6.6% *(14.1)*
5;6-6;3	81.8% (*22.7*)	0% (*0*)	17.6% (*22.7*)

Discussion

Les données issues de l'étude longitudinale confirment la présence précoce de variantes à initiale vocalique. Ainsi que dans l'étude transversale, lors du premier temps d'observation de l'étude longitudinale, les taux des différents types de productions sont semblables (pas de différences significatives avec le test de Wilcoxon). Avec des évolutions internes différentes, on retrouve dans les deux études des tendances qui se rapprochent : la présence constante des séquences avec déterminant, l'augmentation des variantes à voyelle initiale jusqu'à un taux supérieur à 80% et la diminution parallèle des variantes à consonne initiale, qui sont majoritairement en /n/. A partir du temps 5, autour de l'âge de 6 ans, plus aucun enfant de notre échantillon de l'étude longitudinale ne produit les Mots 2 avec une consonne initiale. Il paraît évident, à ce niveau, que les apprentissages alphabétiques jouent un rôle dans les représentations phonologiques. En effet, l'enfant découvre visuellement, vers cette période, la forme écrite à initiale vocalique de ces Mots 2.

3. CONCLUSION

Parmi les questions qui animent le débat sur l'acquisition de la liaison, l'une porte sur le rattachement des consonnes de liaison (CL) au niveau lexical. Il s'agit de savoir si le phonème de liaison est autonome, en coda du Mot 1 ou en attaque du Mot 2 dans les représentations lexicales. Les résultats présentés ici ainsi que ceux de Chevrot & Fayol [5] et Chabanal [6] indiquent que le phonème de liaison pourrait être rattaché, au moins lors de la période de l'acquisition, au Mot 2. Selon Morin [8], la plupart des phonologues situent la CL en finale du Mot 1 sans doute en raison de sa forme écrite. Cet auteur s'oppose à ces conceptions démontrant que la liaison pré-nominale, pouvant être séparée du Mot 1 par une pause, pourrait jouer le rôle de préfixe du Mot 2 (eg : [pəti] *pause* [telefɑ̃]).

Nos données rendent compte de deux éléments :

- il existe une affixation du phonème de liaison à l'initiale du Mot 2 chez l'enfant,

- cette affixation n'est pas systématique (il existe des formes de Mots 2 à initiale vocalique). La thèse exemplariste explique ce type de phénomène en soulignant que l'enfant aurait, pour un même Mot 2, des formes différentes dans son lexique mental, avec et sans phonème de liaison. Précocement, il choisirait alors variablement et aléatoirement un des exemplaires enregistrés. La suite de l'acquisition reposerait sur le fait d'apprendre à associer la bonne "variante" {[nʊʀs], [zʊrs], [tʊʀs], [lʊʀs]} avec le Mot 1 adéquat : *un* + [nʊʀs], *deux* + [zʊʀs], *petit* + [tʊʀs], etc. Or, le processus de connexion qui fonctionne sur un principe de catégorisation permettant la mise en place de la relation morpho-syntaxique reste problématique. En effet, au vu des variations, ce principe semble ne pas aller de soi chez l'enfant qui soit mémorise les contextes sans qu'il y ait eu segmentation, soit découpe la chaîne parlée mais sélectionne un des allomorphes en fonction de sa disponibilité, de sa fréquence et de

facilités articulatoires, sans faire de lien morphologique avec le Mot 1.

Chabanal [6] et Chevrot & Fayol [5] constatent d'ailleurs une forte activation de l'exemplaire en /n/ dans les productions de jeunes enfants. Le phonème /n/ en position initiale de mot étant produit plus tôt dans le développement que le phonème /z/ (Vinter [9]). D'autre part Chabanal [6] observe que les liaisons facultatives les plus souvent produites de manière juste entre 3;4 et 4;2 sont celles qui sont le plus fréquemment utilisées par son entourage. Les contextes "c'est + V" et "il est + V" sont les premiers produits par l'enfant et correspondent aux contextes de liaisons facultatives les plus fréquemment relevés par Boë et Tubach [1].

Il semble donc qu'au départ des principes concrets tels que les questions de facilité articulatoire, de mémorisation, d'effets de fréquence soient plus forts que la mise en place d'un traitement morpho-syntaxique. Il est pourtant au cœur du principe de fonctionnement de la liaison. Le processus de connexion pourrait en réalité fonctionner sur l'identification des Mots 1 comme faisant partie d'une classe sélectionnant telle ou telle CL et sur l'activation dans le lexique mental de l'allomorphe qui convient. L'apprentissage de la liaison se structurerait alors autour de la fréquence des contextes rencontrés. Dans ce cadre, plus l'enfant pourra réunir d'indices de l'existence de points communs par la fréquence des inputs, par exemple l'utilisation de variantes de Mots 2 en /n/ initial après *un*, plus la capacité à inclure de nouveaux éléments dans les ensembles de classe de liaison sera mobilisée. Il s'agit, en d'autres termes, de la capacité à généraliser les connexions "Mot 1-variante de Mot 2" en des schémas plus abstraits (Bybee [4]). Par exemple, si "des" est plus fréquent que "deux", l'apprenti rencontrera plus souvent la séquence "des + CL en /z/" que la séquence "deux + CL en /z/". Il disposera donc de plus d'informations pour rattacher "des" à la classe des unités de gauche qui sélectionnent la variante en /z/. Nous pouvons remarquer des analogies avec la théorie probabiliste des exemplaires. Kirchner [10] postule que les représentations lexicales sont complexes et qu'il existerait différents exemplaires d'un même mot stockés lors des différentes rencontres avec celui-ci. En outre, pour Pierrehumbert [11], le choix de l'exemplaire dépend de sa récence et de sa fréquence dans le lexique mental. Comme l'écrit Dugua [12], "l'influence de la fréquence revient à postuler que plus un exemplaire est souvent activé, plus il sera fort".

Les données obtenues dans cette recherche suggèrent que des formes différentes de Mots 2 co-existent dans le lexique des jeunes enfants. Nos résultats sont par conséquent compatibles avec le fonctionnement lexical impliquant la mise en œuvre de connexions entre items (Chevrot, Dugua & Fayol [13]).

BIBLIOGRAPHIE

[1] L.-J. Boë & J.-P. Tubach. *De A à Zut : dictionnaire phonétique du français parlé*. Grenoble, Ellug, 1992.

[2] P. F. MacNeilage & B. L. Davis. Structure of word forms. *Science*, 288:527-530, 2000.

[3] N. Vallée, I. Rousset & L.-J. Boë. Des lexiques aux syllabes des langues du monde. Typologies, tendances et organisations structurelles. *LINX*, 45:37-50, 2001.

[4] J. Bybee. *Phonology and language use*. Cambridge, Cambridge University Press, 2003.

[5] J.-P. Chevrot & M. Fayol. L'acquisition de la liaison : enjeux théoriques, premiers résultats, perspectives. *Lidil*, 22:11-30, 2000.

[6] D. Chabanal. *Un aspect de l'acquisition du français oral: la variation socio-phonétique chez l'enfant francophone*. Thèse de doctorat en Sciences du langage, Université Paul-Valéry, Montpellier, 2003.

[7] C. Dugua. De la liaison à la formation du lexique chez les jeunes enfants francophones. *Le Langage et l'Homme*, 40, 2:163-182, 2005.

[8] Y.-C. Morin. La liaison relève-t-elle d'une tendance à éviter les hiatus ? Réflexions sur son évolution historique. *Langages*, 158:8-23, 2005.

[9] S. Vinter. Les habiletés phonologiques chez l'enfant de deux ans. *GLOSSA*, 77, 2001.

[10] R. Kirchner. Preliminary thoughts on "phonologization" within an exemplar-based speech processing system. *UCLA Working Papers in Linguistics*, 1:207-231, 1999.

[11] J.-B. Pierrehumbert. Stochastic phonology. *GLOT*, 5, 6:1-13, sous presse.

[12] C. Dugua. *Liaison et segmentation du lexique en français : vers un scénario développemental*. Mémoire de DEA, Université Stendhal, Grenoble, 2002

[13] J.-P Chevrot, C. Dugua & M. Fayol. Liaison et formation des mots en français : un scénario développemental. *Langages*, 158:38-52, 2005.

Changements intonatifs dans la parole Lombard : au-delà de l'étendue de F_0

Pauline Welby

Institut de la Communication Parlée, CNRS UMR 5009, INPG, U. Stendhal, 46 av. F. Viallet, 38031 Grenoble, France
welby@icp.inpg.fr
http://www.icp.inpg.fr/~welby

ABSTRACT

Earlier studies of speech in noise (Lombard speech) have generally reported an increase in fundamental frequency (F_0). This study examined other potential intonational differences. Seven French speakers read a corpus of short paragraphs, in quiet and in 80 dB white noise. Four speakers increased F_0 range in noise. Six upscaled individual tones, although there was great inter-speaker variability. Noise had no effect on intonation pattern type; in particular, there was no tendency to produce more "early rises" in noise, even though these rises are cues to word segmentation. Producing an early rise (thus a LHLH or LHH pattern) may not add to the salience of the commonly produced LH pattern. There were no differences in tonal alignment, in contrast to earlier findings. This null result may be due to paradigm differences between the two experiments.

1. INTRODUCTION

1.1. L'effet Lombard

De nombreuses conversations ont lieu dans du bruit – celui des enfants qui jouent, des voitures qui passent, du vent qui souffle, pour n'en mentionner que quelques-uns. Pourtant, les locuteurs s'adaptent – ils parlent plus fort, modifient la durée des segments de parole (en augmentant la durée des voyelles par rapport à celle des consonnes, en général) et augmentent la fréquence fondamentale (F_0) (voir [8, 10, 14] sur ces modifications et d'autres). On appelle ces modifications l'effet ou le réflexe Lombard. Certaines au moins de ces modifications rendent la parole plus intelligible, aidant ainsi les auditeurs et permettant de transmettre plus efficacement le message [8, 14].

Quelques études ont examiné l'effet du bruit sur le F_0, un effet noté en premier lieu par Lombard lui-même. En général, ces études notent une augmentation de F_0 et une grande variabilité inter-locuteurs (e.g., [8, 14]).

Très peu d'études, néanmoins, ont examiné l'influence du bruit sur les détails de la structure intonative. En fait, en général, les travaux qui ont traité des modifications de F_0 ont utilisé des listes de mots isolés (monosyllabiques pour la plupart), qui ne permettent pas d'examiner la structure intonative. Une étude fait exception, en ayant modélisé l'étendue tonale des cibles de F_0 (la hauteur relative des pics et des creux, le « tonal scaling ») de la parole néerlandaise [12].[1] Dans le bruit, tous les locuteurs ont augmenté leur F_0 pour toutes les cibles. La fonction d'augmentation pourrait être décrite avec un seul modèle pour toutes les cibles, exceptée la dernière cible basse. Un des objectifs de l'étude présentée est d'examiner l'étendue tonale des cibles intonatives du français.

1.2. L'intonation du français

Tous les modèles de l'intonation du français s'accordent sur le fait que l'énoncé est divisé en unités plus petites, appelées Rhythmic Phrase [4], Accentual Phrase (syntagme accentuel, SA) [7], etc. Ce SA est caractérisé par une *montée final*e de F_0 sur sa dernière syllabe (s'il est non final dans l'énoncé) et une *montée initiale* facultative dans sa partie initiale. Des exemples sont donnés sur la figure 1, avec des transcriptions issues de la théorie autosegmentale-métrique, dans laquelle les patrons intonatifs sont considérés comme une série de tons bas ((L)ow) et hauts ((H)igh), les valeurs intermédiaires étant déterminées par interpolation [11]. Outre les patrons LHLH et LLH, un patron possible, souvent observé dans les SA courts, est LH, une montée de L1 à H2 (sans H1 et L2). Le début de la montée initiale (L1) est aligné avec le

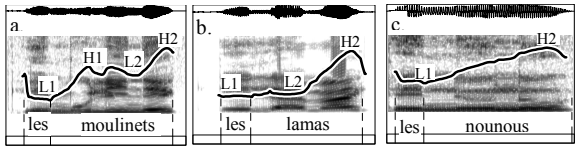


Figure 1 : Des SA avec (a) une montée initiale et une montée finale (LHLH), (b) une montée finale (LLH), (c) une montée d'un L initial à un H final (LH). Exemples tirés de l'étude.

début du premier mot lexical du SA (voir figures 1a, c) [15,17], formant souvent un « coude » dans la courbe de F_0 quand un mot outil précède le mot lexical. On observe ce coude souvent, même en l'absence de montée subséquente. Une montée initiale ou même un simple coude de F_0 (non suivi d'une montée) peuvent être utilisés pour détecter le début d'un mot lexical [16, 17].

1.3. Les changements intonatifs dans la parole Lombard en français

Une etude récente a examiné les modifications articulatoires et intonatives produites par une locutrice française dans le bruit [2, 18]. La parole a été enregistrée dans une condition silencieuse, en présence de bruit « cocktail party » à 85 dB et en présence de bruit blanc à

[1] Plus précisément, les auteurs ne pensent pas que la comparaison porte sur la parole Lombard : « We assume that a considerable portion of the pitch change obtains because speakers deliberately raise their pitch…. We recognize, however, that some portion of the changes will also occur as a by-product of an increase in vocal effort (the "Lombard reflex"…), » mais les motivations pour cette revendication ne sont pas décrites.

85 dB. Trois différences intonatives ont été observées.

Montées initiales

Nous avions émis l'hypothèse selon laquelle les locuteurs produiraient plus de montées initiales dans le bruit, puisque ces montées devraient rendre la parole plus facile à segmenter et donc plus intelligible. Les résultats vont dans le sens de l'hypothèse – la locutrice a produit plus de SA avec des montées initiales dans le bruit. Cependant, il faut interpréter ces résultats avec précaution, d'abord parce qu'ils ne portent que sur une seule locutrice, et ensuite parce que les SA étaient de durée plus longue dans les conditions bruyantes que dans la condition de contrôle silencieuse. Cette différence de durée n'est pas inattendue, étant donné certains résultats de la littérature, mais elle complique l'analyse, puisqu'un SA plus long favorise la réalisation de la montée initiale [15, 17].

L'étendue tonale au sein du SA

La locutrice a augmenté son étendue de F_0 globale de la condition de contrôle silencieuse aux conditions bruyantes. Mais elle a également produit un patron d'étendue tonale inattendu : dans les SA avec un patron à 2 pics (LHLH), la chute du pic de la montée initiale (H1) au début de la montée finale (L2) était beaucoup plus faible dans le bruit que dans le silence. En fait, le L2 avait souvent une valeur de F_0 proche de celle du H1 qui le précédait, comme dans la figure 2b.

Figure 2 : Schéma des différences d'étendue tonale dans des patrons LHLH. (a) chute de H1 à L2, (b) plateau de H1 à L2.

Les patrons des figures 2a et b représentent les extrêmes d'un continuum, pas une dichotomie.[2] Pourtant, les patrons comme celui de la figure 2b sont plus fréquents dans les SA plus courts, pour lesquels les locuteurs n'arrivent pas à atteindre la cible L2 (voir [17], pp. 77–80). Le résultat de [18] est surprenant parce que la locutrice a produit ce patron (« LHplateauLH ») dans certains des SA plus longs.

L'alignement tonal

Les montées initiales produites dans le bruit ne commencent pas toujours à la frontière mot outil/ mot lexical, mais de façon surprenante, parfois au début du mot outil. Cet alignement inhabituel est peut-être dû à l'absence de *feed-back* auditif dans le bruit [5].

1.4. Objectifs de l'étude

L'objectif de cette étude est d'examiner l'influence du bruit sur trois aspects de la structure intonative soulevés par l'étude antérieure : le choix du patron intonatif, l'étendue des tons individuels et l'alignement de ces tons.

2. MÉTHODES

2.1. Corpus

Un corpus de 7 textes courts, chacun contenant 2 à 5 mots cibles, a été construit. Il y avait 11 mots cibles à 2 syllabes et 11 à 3 syllabes. Les cibles ne contenaient que des syllabes CV avec des consonnes sonantes (soit des nasales, soit la latérale /l/). Cette contrainte segmentale a minimisé les perturbations segmentales de la courbe de F_0. Chaque mot cible apparaissait en tête de phrase, précédé d'un article (*le, la, les* ou *un*). Ce syntagme était suivi d'une proposition relative commençant par un pronom relatif (*qui, que, qu'il*), suivie du syntagme verbal de la proposition principale. Un exemple avec cibles soulignées est présenté ci-dessous (1).

(1) Il y eut une scène chaotique à la crèche cet après-midi. Un mulot qui s'était réfugié au fond de la cantine avait fait peur aux petits. Ils s'étaient tous mis sur leurs chaises en hurlant. Les moulinets que Daniel était en train de faire avec les bras avaient fait tomber le bocal à poissons. Il y avait de l'eau et du verre partout. Heureusement, Charlotte pensa à mettre le pauvre petit poisson rouge dans un verre d'eau. Les nounous qui étaient venues chercher les enfants secouaient la tête en regardant la scène.

Chaque participante était assise à une table en face de la chercheuse et reçut la consigne « lisez les paragraphes comme si vous les lisiez à la personne en face de vous ».

2.2. Participantes

Sept locutrices natives du français hexagonal ont participé à l'étude. La Locutrice 7 était la locutrice de [18].

2.3. Procédures

Les participantes ont lu le corpus 4 fois, d'abord dans le silence avec un débit de parole normal puis avec un débit rapide, ensuite dans du bruit avec les deux débits. Du bruit blanc a été diffusé par écouteurs à 80 dB. Les locutrices ont été enregistrées avec un microphone-casque et un enregistreur numérique à 22,05 kHz, et les données ont été transférées sur ordinateur.

Les deux débits ont été utilisés uniquement pour obtenir de la parole avec des durées comparables pour les conditions de silence et de bruit, afin d'éviter la complication de l'analyse de l'étude précédente [18]. Nous espérions que la tendance à ralentir en lisant dans le bruit serait mitigée par le fait que la condition à débit normal bruyante suivait directement la condition à débit rapide silencieuse.

Les fichiers audio ont été segmentés et chaque énoncé sauvegardé comme un fichier individuel. Les courbes de F_0 et les spectrogrammes ont été créés sous Praat [3]. Les frontières de mots ont été étiquetées, en s'appuyant sur le signal et le spectrogramme, et à l'aide de scripts Praat qui ont permis de semi-automatiser le processus. (Pour une description de la procédure d'étiquetage, voir [15, 17], http://www.icp.inpg.fr/~welby/praat.html.)

3. RÉSULTATS

3.1. Variation de débit

De façon générale, la manipulation du débit n'a pas permis de minimiser les différences de durée entre la condition à débit normal silencieuse et la condition à débit normal bruyante. Une seule locutrice, la Locutrice 3, n'a pas produit de différence de durée significative entre ces

[2] Cette répartition va à l'encontre des modèles dans lesquels ces 2 contours (avec soit une chute de H1 à L2, soit un plateau) sont deux éléments distincts dans un inventaire de patrons holistiques (e.g., [13]).

deux conditions ($t(21) = 0{,}110$; $p = 0{,}914$). Une deuxième locutrice (la Locutrice 6) a des SA plus longs dans la condition à débit normal bruyante que dans la condition à débit normal silencieuse ($t(21) = -4{,}379$; $p < 0{,}001$). Les autres locutrices ont des durées plus courtes dans la condition à débit normal bruyante (toutes les valeurs de $p < 0{,}01$, à l'exception de la Locutrice 4, $p < 0{,}05$).

3.2. Patrons intonatifs observés

Contrairement aux prédictions, aucune locutrice n'a augmenté significativement le taux de montées initiales de la condition silencieuse à la condition bruyante.

Presque 90 % des mots cibles à 3 syllabes ont été produits avec une montée initiale, la plupart avec le patron LHLH, quelques-uns avec le patron LHH.

Pour les mots cibles à 2 syllabes, on a observé plus de variété dans les patrons intonatifs. Seuls 15 % des SA dans les conditions à débit normal silencieuse et à débit normal bruyante ont été produits avec soit LHLH soit LHH, tandis que 84 % ont été produits avec soit LH soit LLH. Trois locutrices sur 7 ont semblé préférer le patron LH, mais 3 locutrices ont produit de nombreux patrons LLH. La Locutrice 7 a produit de nombreux patrons LLH, mais aussi quelques patrons LHLH.

3.3. Alignement tonal

Nous avons inspecté visuellement l'alignement tonal du début de la montée initiale (L1) dans les conditions à débit normal silencieuse et bruyante. Les SA courts avec le patron LH n'ont souvent pas de coude, et la montée peut commencer tout au début du SA. Des SA plus longs avec

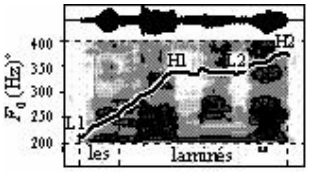


Figure 3 : Le SA *les laminés* réalisé avec le patron LHLH. (de la condition à débit normal bruyante de la Locutrice 6).

le patron LHLH, cependant, ont typiquement deL s coudes L1. Contrairement à l'étude précédente, nous n'avons pas observé pour ces SA un alignement de L1 moins précis dans le bruit ; la plupart de ces SA a des coudes L1 à la frontière mot outil/ mot lexical. Néanmoins, nous avons observé quelques SA LHLH dans lesquels L1 était réalisé au début du mot outil (voir la figure 3).

3.4. Étendue tonale

L'étendue de F_0 sur le SA cible (H2 – 1) a été examinée ; quatre locutrices ont augmenté leur étendue de F_0 du silence au bruit, comme le montre la figure 4.[3] Des analyses de variance à 1-facteur ont été réalisées pour

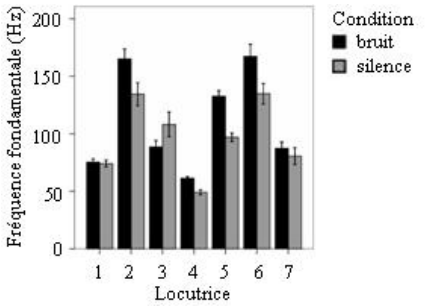


Figure 4 : Étendue de F_0 dans des SA cibles par condition de bruit. Les barres indiquent l'erreur standard de la moyenne.

déterminer si la condition de bruit a influencé l'étendue des 4 tons du patron LHLH, le plus fréquent des patrons. Les résultats sont donnés dans la table 1.[4]

Table 1 : Valeurs de F_0 pour les 4 tons du patron LHLH par condition. Les différences significatives sont marquées par '*'.

S	L1		H1		L2		H2	
	sil.	bruit	sil.	bruit	sil.	bruit	sil.	bruit
1	197	198	253	257	235	249	265	281
2	203	217*	257	272	251	290*	316	393*
3	193	199	270	260	214	216	279	274
5	208	227*	304	337*	260	292*	309	355*
6	216	231	305	331*	272	296*	334	373
7	179	186*	248	242	201	221*	255	265

Seules les Locutrices 2 et 7 (la locutrice dans [18]) ont produit des patrons proches du « LHplateauLH » observé dans [18]. Dans le bruit, elles ont augmenté L2 mais pas H1, ce qui a conduit à une plus petite chute de H1 à L2.

4. DISCUSSION ET CONCLUSIONS

Parmi les paramètres examinés, les résultats se sont avérés les plus fructueux pour l'étendue tonale. Pour toutes les locutrices sauf une, nous avons constaté des différences significatives dans l'étendue des tons individuels entre les conditions de silence et de bruit, même s'il y avait une variabilité inter-locutrices importante en ce qui concerne quel(s) ton(s) étai(en)t impliqué(s). Cette variabilité n'est pas surprenante : certains travaux sur le français ont montré de grandes différences inter-locuteurs pour l'étendue [6, 19]. Le fait que l'on trouve des différences d'étendue des tons individuels est d'autant plus remarquable que toutes les locutrices n'ont pas augmenté leur étendue de F_0 globale dans la condition bruyante (un résultat qui s'accorde avec la variabilité observée dans la littérature). Une mesure de l'étendue globale de F_0 n'est donc pas adaptée à ce type de différence.

Les différence entre les résultats de notre étude, dans laquelle 7 locutrices françaises ont montré une grande variabilité dans l'étendue des tons individuels, et ceux de

[3] Comme un des relecteurs l'a suggéré, il serait intéressant de voir si l'augmentation de F_0 est corrélée avec l'augmentation d'intensité typiquement observée dans le bruit. Malheureusement, une telle mesure n'est pas possible pour ce corpus, la distance microphone-bouche n'étant pas constante. (La mise en place des écouteurs pour la condition de bruit a souvent nécessité un repositionnement du microphone.) Nous remarquons, néanmoins, qu'une augmentation de F_0 n'est pas systématiquement observée, ni pour les locutrices de cette étude, ni dans d'autres études (contrairement à une augmentation d'intensité).

[4] Les données de la Locutrice 4 ont été exclues, puisqu'elle a produit trop peu (3) de patrons LHLH dans la condition bruyante.

[12], dans laquelle 15 locuteurs néerlandais ont systématiquement déplacé tous les tons vers le haut, sont peut-être linguo-spécifiques. L'étendue tonale varie d'une langue à une autre (e.g., le français vs. le néerlandais et le grec [1]) pour la parole produite dans des conditions claires. C'est peut-être aussi le cas de la parole Lombard. Puisque la forme exacte du F_0 du SA ne signale pas, en général, une différence de sens en français (au moins pour les SA non ambigus), il semble possible que les locuteurs n'aient pas besoin de préserver toutes les caractéristiques de l'étendue tonale.

Il est aussi possible que ces différences soient dues à des différences de paradigmes expérimentaux, puisque des facteurs telles que l'intensité et le type de bruit pourraient influencer l'étendue tonale.

L'étude présentée ici n'a pas montré un alignement moins précis de L1 dans le bruit, même pour la Locutrice 7 (la locutrice dans [18]). Ce résultat contraire à celui de [18] est peut-être dû à des différences de paradigmes expérimentaux. L'étude antérieure a utilisé du bruit à 85 dB diffusé par une enceinte (ce qui a mis la locutrice comme les chercheuses/ auditrices dans un environnement bruyant), alors que dans l'étude présentée ici, le bruit était à 80 dB et diffusé par des écouteurs. Il est possible que l'alignement tonal ne soit perturbé qu'au-delà d'un certain seuil ou avec un certain degré d'effort vocal.

Finalement, les locutrices n'ont pas produit plus de montées initiales dans le bruit. Ce résultat est peut-être dû au fait que pour la plupart des locutrices, les SA ont été produits avec une durée moins longue dans le bruit, ce qui défavorise la réalisation de la montée initiale. Cependant, nous n'avons pas constaté d'augmentation de taux de montées initiales, même pour les deux locutrices dont les SA n'étaient pas de durée plus courte dans les conditions bruyantes. Il se pourrait que la montée du patron LH, que l'on observe souvent dans des SA courts, soit aussi efficace comme indice de la segmentation lexicale que la montée initiale du LHLH. Plus précisément, alors que l'ajout d'une montée initiale à un patron LLH (donnant LHLH) pourrait rendre plus saillant le début d'un mot lexical, il se pourrait qu'il n'existe aucun avantage pour LHLH versus LH.

De nombreux auteurs ont noté l'influence de la tâche dans l'étude de la parole Lombard (e.g., [9]). En particulier, augmenter la charge communicative (en utilisant des paires locuteur-auditeur, par exemple) pourrait conduire à une accentuation de certains aspects de l'effet.

La plupart des études Lombard ont supposé au moins tacitement que certains des changements sont indépendants de la langue considérée, comme le terme « *réflexe* Lombard » le suggère. Même si l'on ne peut pas conclure des résultats de notre étude qu'il existe des indices linguo-spécifiques, cette question mérite d'être examinée.

BIBLIOGRAPHIE

[1] A. Arvaniti, D. R. Ladd et I. Mennen. Stability of tonal alignment: the case of Greek prenuclear accents. *J. Phon.,* 26 : 3–25, 1998.

[2] L. Bailly, *Étude articulatoire de la parole produite en environnement bruyant*. Mémoire de master. Laboratoire d'Acoustique Musicale, Université de Paris 6, 2005. http://www.www.icp.inpg.fr/~lbailly/RapportFinal_LucieBailly.pdf

[3] P. Boersma et D. Weenink. PRAAT: Doing phonetics by computer (version 4.3.28), 2005. http://www.praat.org

[4] A. Di Cristo. Intonation in French. Dans D. Hirst et A. Di Cristo (éds.), *Intonation Systems: a survey of twenty languages*, pages 195–218. Cambridge : Cambridge University Press, 1998.

[5] L. Elliott et A. Niemoeller. The role of hearing in controlling voice fundamental frequency. *Intl. Audiology,* 9 : 47–52, 1970.

[6] C. Fougeron et S.-A. Jun. Rate effects on French intonation: Prosodic organization and phonetic realization. *J. Phon.,* 26 : 45–69, 1998.

[7] S.-A. Jun et C. Fougeron. Realizations of accentual phrase in French. *Probus,* 14 : 147–172, 2002.

[8] J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *JASA,* 93 : 510–524, 1993.

[9] H. Lane et T. Bernard. The Lombard sign and the role of hearing in speech. *J. Speech & Hearing Research,* 14 : 677–709, 1971.

[10] É. Lombard. Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx*, 37 : 101–119, 1911.

[11] J. B. Pierrehumbert. *The phonology and phonetics of English intonation*. Thèse de doctorat, MIT, 1980.

[12] E. Shriberg, D. R. Ladd, J. Terken et A. Stolcke. Modeling pitch range variation within and across speakers: Predicting F0 targets when "speaking up." *Proc. ICSLP 4*, pages 1–4 (annexe des actes), 1996. Version corrigée disponible à http://www.asel.udel.edu/icslp/cdrom/vol2/553/a553.pdf

[13] J. Vaissière. Une procédure de segmentation automatique de la parole en mots prosodiques en français. *Actes des VIIèmes JEP,* pages 103–114, 1977.

[14] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow et M. A. Stokes. Effects of noise on speech production: acoustic and perceptual analyses. *JASA,* 84 : 917–928, 1988.

[15] P. Welby, French intonational structure: evidence from tonal alignment. *J. Phon.,* sous presse.

[16] P. Welby. French intonational rises and their role in speech seg mentation. *Proc. Eurospeech 8*, pages 2125–2128, 2003.

[17] P. Welby. *The slaying of Lady Mondegreen, being a study of French tonal alignment and association and their role in speech segmentation*. Thèse de doctorat, The Ohio State University, 2003. http://www.ling.ohio-state.edu/publications/dissertations

[18] P. Welby, L. Bailly, M. Garnier, M. Dohen et H. Lœvenbruck. Pitch changes in noisy conditions: data from French. Dans *Proc. Making Europe More Attractive for Researchers,* à par.

[19] P. Welby et H. Lœvenbruck. Anchored down in Anchorage: syllable structure and segmental anchoring in French. *Italian Journal of Linguistics. Special issue on Autosegmental-metrical approaches to intonation in Europe: Tonal targets and anchors,* M. D'Imperio (éd), à par.

REMERCIEMENTS

Je tiens à remercier Hélène Lœvenbruck pour ses conseils précieux et pour son aide avec la construction du corpus, et Denis Beautemps, Xavier Pelorson et Coriandre Vilain pour leur assistance technique. Je remercie également les participantes. Cette recherche a été soutenue par une bourse internationale Marie Curie du 6th European Community Framework Programme.

Le paradigme ascendant de la FO dans les fonctions « préindicatives » adverbiales en portugais brésilien

Cirineu Cecote Stein

Universidade Federal do Rio de Janeiro,
Laboratório de Fonética Acústica / CAPES
Rio de Janeiro, Brasil
cirineustein@uol.com.br

ABSTRACT

This paper presents the results of two perception tests, which aimed to identify, in Brazilian Portuguese, if a native user of the language, not trained in phonetics, is able to recognize a pre-indicative adverbial value, and, with the help of vocal synthesis techniques, which specifications of the prosodic components are responsible for the establishment of that value. Although the whole of the research focuses on nine adverbial semantic fields, only three of them will be discussed here. These three possible pre-indicative patterns of cause, consequence and finality show an ascending melodic curve in the final stressed vowel. Due to the similarity among the melodic curve contours in these three pre-indicative patterns, it is possible that a careless listener perceives them as ambiguous.

1. INTRODUCTION

L'élocution d'un énoncé peut laisser transparaître plusieurs intentions de l'énonciateur, telles qu'une attitude ou une émotion (Léon [6]). Fónagy [2] a caractérisé ce que l'on connaît comme la fonction « préindicative » de l'intonation : à partir de l'élocution d'une proposition, il est possible de prévoir, dans certains cas, ce qui sera dit dans la suite. Plusieurs chercheurs se sont alors penchés sur la relation syntaxe-prosodie (Fox [3], Ladd [5]). Pour le portugais brésilien, on peut citer Freitas [4]. Toutefois, à notre connaissance, il n'y a pas d'études spécifiques sur le sujet de cet article. Dans le groupe des propositions adverbiales, neuf catégories sont traditionnellement retenues dans le portugais brésilien [1] : cause, conséquence, condition, conformité, concession, comparaison, but, proportionnalité et temporalité. Si l'on considère que chacune de ces catégories présente un champ sémantique spécifique, on s'attendra à l'existence des réalisations prosodiques distinctes pour chacune d'elles dans les propositions principales, si ces propositions principales n'offrent aucune détermination sémantique. Ces réalisations prosodiques sont considérées comme la fonction préindicative de l'intonation, selon la terminologie adoptée par Fónagy [2].

Cet article met en évidence les résultats obtenus, à partir de deux tests perceptifs, pour la caractérisation de ce que l'on dénommera le « paradigme ascendant de la FO », dans trois parmi neuf catégories de propositions adverbiales en portugais brésilien : cause, conséquence et but. Ces manifestations pourraient présenter une certaine ambigüité à un auditeur peu attentif, à cause de la ressemblance entre les éléments prosodiques. Cette apparente ambigüité se résoud dans le comportement particulier de la courbe mélodique et de la durée de quelques segments dans certains points critiques de la proposition principale.

Pour le second test, des synthèses de deux propositions principales monotoniques ont été produites, comme tentative de rapprocher le contour de la FO de ce que l'on observe dans les propositions principales originalement utilisées dans le premier test. Au-delà de la FO, la durée de quelques segments, considérés critiques, a été manipulée.

2. TEST 1

2.1. Méthodologie

Le premier test réalisé a eu pour objectif de repérer si l'usager courant de la langue, sans entraînement spécifique en phonétique, est capable de reconnaître l'existence, dans les propositions principales, de la préindication prosodique des champs sémantiques présents dans les propositions adverbiales. En raison de la subtilité extrême de cette perception, on a choisi un test à choix binaire, dans lequel toutes les neuf possibilités de préindication adverbiale seraient présentes.

Dix-huit propositions ont été enregistrées par une parleuse brésilienne native, expérimentée en phonétique. Neuf de ces propositions ont été introduites par une proposition principale PP1 – *ficava infeliz* [il (ou elle) devenait triste] – et neuf autres par une proposition principale PP2 – *mostrava-se cansado* [il (ou elle) avait l'air fatigué(e)]. Les propositions adverbiales qui ont suivi ces propositions principales étaient introduites par un connecteur qui explicitait le champ sémantique suggéré : cause, conséquence, condition, conformité, concession, comparaison, but, proportionnalité et temporalité. Les mêmes propositions adverbiales ont été utilisées aussi bien pour la PP1 que pour la PP2. Pour les 18 propositions enregistrées, on a relevé les propositions principales, que l'on nommera désormais PP, en produisant des fichiers sonores indépendants.

Le test a été appliqué à 41 étudiants volontaires en 1^er et 2^e cycles de la Faculté de Lettres de l'Universidade Federal do Rio de Janeiro, qui ne connaissaient pas les objectifs

du test. Pour chaque séquence, deux PPs, aux valeurs préindicatives distinctes, dénommées (A) et (B), ont eté présentées, par écrit, à côté de deux propositions adverbiales distinctes, dénommées (1) et (2). Entre les PPs et les propositions adverbiales il y avait un espace, où le participant devrait indiquer le croisement de sa préférence, par exemple, A2 B1. Étant donné que le contenu sémantique des deux PPs présentes dans la séquence était le même, il fallait choisir sur la base de la modulation prosodique entendue pour chaque PP. Pour que toutes les possibilités de croisement, par analyse combinatoire, soient testées, 72 séquences ont été produites. Chaque séquence sonore était constituée de deux PPs impliquées, répétées trois fois, avec un intervalle de 400 ms entre les PPs de la paire et de 700 ms entre chaque paire. En cas de besoin, les participants pourraient solliciter une nouvelle audition immédiate de la séquence. Une nouvelle séquence sonore était présentée seulement après que tous avaient choisi. L'application du test a été divisée en quatre séances, au cours d'une même semaine.

2.2. Résultat

Le résultat du test a mis en évidence que, parmi les neuf paradigmes préindicatifs possibles des propositions adverbiales, six sont bien reconnus par l'usager courant de la langue. L'objectif était de vérifier avec lesquels des autres huits paradigmes un certain paradigme ne se confondrait pas. Pour le calcul du résultat final, on a considéré, premièrement, comme suffisamment reconnaissable le paradigme pour lequel la signifiance était égale ou inférieur à 0,05, selon un test chi-carré. Deuxièmement, parmi les huit possibilités de contraste (cause en contraste avec comparaison, concession etc.), au moins cinq d'entre elles devraient être repérées par au moins les deux tiers des participants.

A partir de ces critères, il a été possible de constater que l'usager courant de la langue tend à reconnaître plus facilement les PPs qui préindiquent des propositions adverbiales causales, concessives, conformatives, consécutives, temporelles et finales, selon les caractéristiques prosodiques présentes dans les PPs utilisées pour le teste 1. L'analyse visuelle de la courbe mélodique de ces PPs démontre que trois parmi elles – cause, conséquence et but – présentent la FO avec une forme ascendante sur la voyelle de la syllabe tonique finale, comme le montre la figure 1.

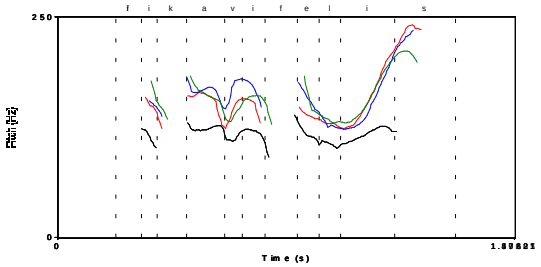


Figure 1 : courbe mélodique de la PP1 – *ficava infeliz*, prononcée d'une façon monotonique (en noir) et avec des préindications causale (en rouge), consécutive (en bleu) et finale (en vert)

3. SYNTHÈSES

Une fois repérées les manifestations prosodiques des PPs qui constituent possiblement les paradigmes prosodiques préindicatifs des propositions adverbiales, dans le portugais brésilien, on a eu recours au processus de synthèse de la parole, comme une manière d'identifier dans quelle mesure les éléments prosodiques influencent dans la caractérisation et l'identification de chacun des paradigmes préindicatifs. En observant la figure 1, on perçoit la subtilité extrême de la différence entre les trois manifestations dont la fréquence fondamentale est ascendante.

Un phonéticien brésilien expérimenté a enregistré, pour les mêmes PPs 1 et 2, une version tendant à la monotonicité. Le processus de synthèse, réalisé avec le logiciel Praat v.4.2.29, a consisté à adapter les éléments prosodiques de ces deux versions monotoniques à la conformation vérifiée dans les originaux sonores de la PP1 enregistrés pour le test 1, dénommés ici « originaux ». En considérant les trois paradigmes de la FO ascendante sur la voyelle de la syllabe tonique finale, le procédé est le suivant :

3.1. préindication consécutive

3.1.1. centralisation du sommet formé par la FO sur la première consonne de la syllabe tonique finale ;

3.1.2. la FO de la voyelle de la dernière syllabe tonique, dans l'original, présente une forme ascendante, diamétralement opposée à la courbe descendante formée par la voyelle de la syllabe prétonique et la consonne de la syllabe tonique finale. On considère une ligne de base parallèle à l'axe temporel du graphique, qui passe par le pic de la voyelle de la syllabe prétonique finale. La variation de la FO entre cette ligne de base et le point le plus bas de la consonne de la syllabe tonique finale est de 30%, indice que l'on va considérer comme « valeur de réference » (VR). Au-delà de cette ligne de base, transposée à la PP monotonique, la FO de sa voyelle, dans la syllabe tonique finale, a été élevée :
   a) en 1 VR ;
   b) avec 25% d'augmentation de la VR ;
   c) avec 50% d'augmentation de la VR ;
   d) avec 75% d'augmentation de la VR ;
   e) avec 100% d'augmentation de la VR ;
   f) avec 125% d'augmentation de la VR ;
   g) avec 150% d'augmentation de la VR ;
   h) avec 175% d'augmentation de la VR ;
   i) avec 200% d'augmentation de la VR ;

3.1.3. cumulativement, il y a une augmentation de la durée de la consonne dans la syllabe tonique

finale en 91,5%, similaire à la durée vérifiée à l'original.

L'audition, par l'auteur, des synthèses produites a indiqué que la transformation (3.1.1) n'influence pas le résultat final, et que, pour les transformations réalisées dans (3.1.2), le point critique est l'item (f) : en dessous de ce coefficient, on n'obtient pas la préindication consécutive ; en dessus de ce coefficient, l'accent est mis sur la préindication. La transformation (3.1.3) rend la totalité de la synthèse plus proche de l'original et, appliquée aux items de (a) à (e) de (3.1.2), elle n'est pas capable de leur conférer la valeur consécutive.

3.2. préindication causale

3.2.1. pour la manipulation de la FO de la voyelle de la syllable tonique finale, on a adopté comme référence la transformation signalée à (3.1.2.f) ;

3.2.2. cumulativement, l'élévation de toute la FO antérieure à la syllabe prétonique
   a) de 10% ;
   b) de 20%.

L'audition, par l'auteur, des synthèses produites a indiqué que la transformation (3.2.1) paraît déjà suffire pour conférer à la PP monotonique la valeur préindicative causale, et que la transformation (b), à (3.2.2.), cumulative à (3.2.1), rend le résultat plus performant.

3.3. préindication finale

3.3.1. visuellement, ce qui paraît distinguer le paradigme préindicatif final du paradigme causal, dans la PP1, c'est le point maximum d'élévation de la FO dans la voyelle de la syllabe tonique finale. Dans les originaux, la FO de la préindication causale atteint un point 30% supérieur à celui de la finale, autrement dit, tandis que la FO de la préindication causale s'élève avec une variation de 92% en dessus de la ligne de base, celle de la préindication finale s'élève de 62%. La ligne de base, dans ce cas, a été considérée comme le point initial d'élévation de la courbe, en Hz, situé sur la consonne de cette syllabe tonique finale. Application de cette transformation ;

3.3.2. modulation de la FO, dans la voyelle de la syllabe tonique finale, en lui conférant une forme à 5% de concavité ;

3.3.3. la durée de la consonne de la syllabe tonique finale, dans l'original, est 40% supérieure que pour la PP monotonique. Application de cette transformation ;

3.3.4. la durée de la syllabe /ka/, dans la PP1 original, correspond à 72% de la durée de cette syllabe dans la PP1 monotonique. Réduction de la durée de la syllabe /ka/, dans la PP monotonique donc de 28% ;

3.3.5. la durée de la voyelle et de la consonne dans la syllabe tonique finale est inférieure dans l'original à celle de la PP1 monotonique, respectivemente 37% et 91%. Application de cette transformation.

L'audition, par l'auteur, des synthèses produites a indiqué qu'il paraît nécessaire que la transformation (3.3.1) soit accompagnée de (3.3.3). La transformation (3.3.2) est dispensable. Les transformations (3.3.4) et (3.3.5), cumulativement à (3.3.1) et (3.3.3), rendent les synthèses plus proches de l'original, alors qu'elles ne semblent pas être indispensables.

4. TEST 2

4.1. Méthodologie

L'objectif du test 2 a été de vérifier quelles transformations produites par les synthèses sont responsables de la valeur préindicative des PPs. Comme il s'agissait du jugement des transformations dans une proposition monotonique, on a préféré un modèle dont une seule PP était analysée chaque fois. Pour ce test, ont été utilisés seulement les six paradigmes préindicatifs au plus haut indice de reconnaissance dans le test 1 et les synthèses qui, selon la perception personnelle de l'auteur, constitueraient plus probablement des points critiques. Pour la PP1 comme pour la PP2, ont étés sélectionnées, pour la préindication consécutive, les synthèses (3.1.2.f) et (3.1.3) ; pour la préindication causale, les synthèses (3.2.1) et (3.2.2.b) ; et pour la préindication finale, les synthèses (3.3.1) et (3.3.3), dont les courbes mélodiques peuvent être visualisées dans les figures 2, 3 et 4, ci-dessous.

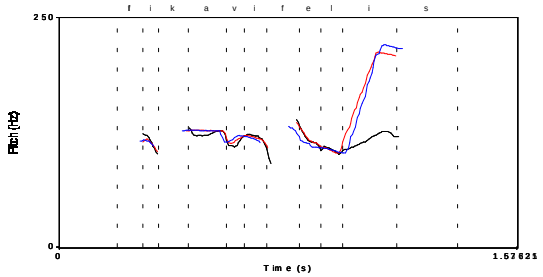


Figure 2 : courbes mélodiques des synthèses produites depuis la PP1 monotonique – *ficava infeliz* – (en noir), pour la préindication consécutive. Synthèse (3.1.2.f) en rouge et synthèse (3.1.3) en bleu

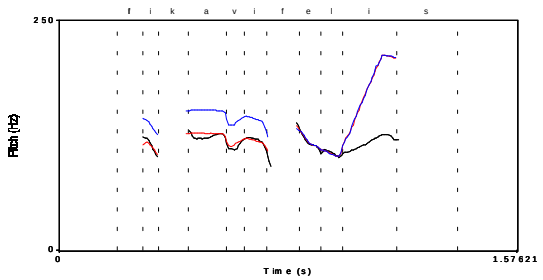


Figure 3 : courbes mélodiques des synthèses produites depuis la PP1 monotonique – *ficava infeliz* (en noir), pour

la préindication causale. Synthèse (3.2.1) en rouge et synthèse (3.2.2.b) en bleu

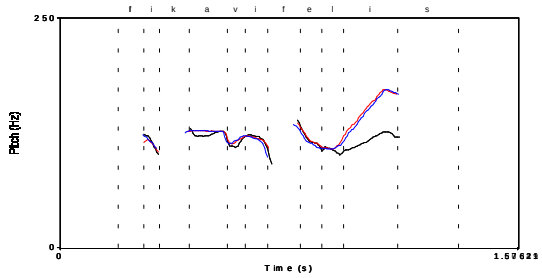


Figure 4 : courbes mélodiques des synthèses produites depuis la PP1 monotonique – *ficava infeliz* (en noir), pour la préindication finale. Synthèse (3.3.1) en rouge et synthèse (3.3.3) en bleu

Chaque séquence a présenté une PP dans la première colonne, suivie, dans une seconde colonne, par deux propositions adverbiales, précédées de parenthèses. Les 28 participants, des étudiants volontaires de 1er et 2e cycles de la Faculté de Lettres de l'Universidade Federal do Rio de Janeiro, devraient marquer l'une des deux propositions subordonnées qui compléterait le mieux la proposition principale, selon l'enregistrement écouté. Chaque PP a été répétée trois fois, avec un intervalle de 2300 ms entre chaque écoute. Les participants n'ont pas été informés de l'objectif du test. Pour que toutes les possibilités de croisement, par analyse combinatoire, soient testées, 120 séquences ont été produites. Le test 2, de même que le test 1, a été divisé en quatre séances au cours d'une même semaine.

4.2. Résultats

Les résultats obtenus pour les synthèses relatives aux paradigmes préindicatifs avec la FO ascendante dans la voyelle tonique finale ont mis en évidence que :

a) pour la valeur préindicative causale, l'indice de reconnaissance pour la synthèse (3.2.2.b) a été de 13,5% supérieur à celui pour la synthèse (3.2.1) ;
b) pour la valeur préindicative consécutive, l'indice de reconnaissance pour la synthèse (3.1.3) a été de 14% supérieur à celui de la synthèse (3.1.2.f) ;
c) pour la valeur préindicative finale, l'indice de reconnaissance pour la synthèse (3.3.3) a été de 15,8% supérieur à celui de la synthèse (3.3.1).

5. CONCLUSION

Selon les résultats obtenus, il semble possible d'établir, pour l'effet de la synthèse de la parole, que, pour suggérer des valeurs prosodiques préindicatives des propositions adverbiales causales, consécutives et finales, la proposition principale monotonique devra être modulée avec les caractéristiques suivantes :

a) valeur préindicative causale: la FO de la voyelle de la syllabe tonique devra présenter une forme ascendante, en s'élevant de 67,5% à partir de la ligne de base, caractérisée en (3.1.2), ci-dessus ;
b) valeur préindicative consécutive: la FO de la voyelle de la syllabe tonique devra présenter une forme ascendante, en s'élevant de 67,5% à partir de la ligne de base, caractérisée en (3.1.2), ci-dessus. Cumulativement, l'augmentation de la durée de la consonne dans la syllabe tonique finale est de 91,5% ;
c) valeur préindicative finale: l'élévation de la FO dans la syllabe tonique finale est de 62% par rapport à la ligne de base, considérée, dans ce cas, comme le point initial de l'élévation de la courbe, placé sur la consonne de cette syllabe tonique finale. Cumulativement, l'augmentation de la durée de la consonne de la syllabe tonique finale est de 40%.

Je remercie le CAPES pour son soutien (processus BEX4431/05-7) et les deux relecteurs pour leurs précieux commentaires.

BIBLIOGRAPHIE

[1] C. Cunha, L. Cintra. *Nova gramática do português contemporâneo.* Nova Fronteira, Rio de Janeiro, 1985.

[2] I. Fónagy. *La fonction préindicative de l'intonation en français et en hongrois.* In *Travaux de l'Institut d'Études Linguistiques et Phonetiques*, volume 1, pages 44-75, 1974.

[3] A. Fox. Subordinating and Co-ordinating Intonation Structures in the Articulation of Discourse. In *Intonation, Accent and Rythm Studies – Phonology.* Daffyd Gibbon and Helmut Richter, eds. Walter de Gruyter, Berlin, 1984.

[4] M.A. de Freitas. *Prosódia e sintaxe: delimitação e contraste de estruturas.* Thèse de doctorat. UFRJ, Rio de Janeiro, 228 pages, mimeo., 1995.

[5] D.R. Ladd Jr. Intonation, main clause phenomena, and point of view. In L.R Waugh & C.H. Van Schooneveld. *The melody of language. Intonation and prosody*. University Park Press, Baltimore, pages 149-163, 1980.

[6] Léon, P. De l'analyse psychologique à la catégorisation auditive et acoustique des émotions dans la parole. *Journal de Psychologie*, vol. 3-4, pages 305-323.

Session X

Conférence Invitée

Mercredi 14 juin 2006 - 09h00 10h00

Open Domain Speech Translation: From Seminars and Speeches to Lectures

Christian Fügen[*], *Muntsin Kolss*[*], *Matthias Paulik*[†], *Sebastian Stüker*[*],
Tanja Schultz[†], *Alex Waibel*[*†]

[*]Interactive Systems Labs (ISL)
Universität Karlsruhe (TH), Germany
{fuegen, kolss, stueker, waibel}@ira.uka.de
[†]Interactive Systems Labs (ISL)
Carnegie Mellon University, Pittsburgh, PA, USA
{paulik, tanja}@cs.cmu.edu

ABSTRACT

This paper describes our ongoing work in domain unlimited speech translation. We describe how we developed a lecture translation system by moving from speech translation of European Parliament Plenary Sessions and seminar talks to the open domain of lectures. We started with our speech recognition (ASR) and statistical machine translation (SMT) 2006 evaluation systems developed within the framework of TC-Star (Technology and Corpora for Speech to Speech Translation) and CHIL (Computers in the Human Interaction Loop). The paper presents the speech translation performance of these systems on lectures and gives an overview of our final real-time lecture translation system.

1. INTRODUCTION

Growing international information structures and decreasing travel costs could make the dissemination of knowledge in this globalized world very easy – if only the language barrier could be overcome. Lectures are a very effective method of knowledge dissemination. Such personalized talks are the prefered method since they allow the speakers to tailor their presentation toward a specific audience, and in return allow the listeners to get the most relevant information through interaction with the speaker. In addition, personal communication fosters the exchange of ideas, allows for collaboration, and forms ties between distant units, e.g. scientific laboratories or companies. At the same time it is desirable to allow the presenters of talks and lectures to speak in their native language, since, no matter how proficient in a foreign language, one will always feel more confident in the native tongue. To overcome this obstacle human translators are currently the only solution. Unfortunately, translation service are often prohibitively expensive such that many lectures are not given at all as a result of the language barrier. The use of modern machine translation techniques have the potential to provide translation services at no costs to a wide audience, making it possible to overcome the language barrier and bring the people closer together.
This paper describes our ongoing work in unlimited domain speech translation of lectures starting from systems built within the framework of CHIL and TC-STAR.

CHIL [25], *Computers in the Human Interaction Loop*, aims at making significant advances in the fields of speaker localization and tracking, speech activity detection and distant-talking automatic speech recognition. Therefore, in addition to the near and far-field microphone, seminars were also recorded by calibrated video cameras. The long-term goal is the ability to recognize speech in a real reverberant environment, without any constraint on the number or distribution of microphones in the room nor on the number of sound sources active at the same time.

TC-STAR [20], *Technologies and Corpora for Speech-to-Speech-Translation*, is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST) which is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS). The objective of the project is to make a breakthrough in SST that significantly reduces the gap between human and machine translation performance. The focus hereby is on the development of new algorithms and methods. So far the project targets a selection of unconstrained conversational speech domains – speeches and broadcast news – and three languages: European English, European Spanish, and Mandarin Chinese.

The paper is organized as follows: The developmental work started from our 2006 ASR and SMT evaluation systems for European Parliament Plenary Session (EPPS, TC-STAR) and the NIST Rich Transcription evaluation RT-06S on seminars (CHIL). In Section 3, we first compare the different ASR systems of both domains and show how we merged these systems for lecture recognition. Furthermore, we present first results of acoustic and language model adaptation on the lecture domain. In Section 4, we give statistical machine translation results on text and ASR input for lectures of our 2006 SMT evaluation system for EPPS. In addition, we explain in detail how we adapted our EPPS SMT system towards the more conversational style of lectures and present the corresponding machine translation results. Section 5 provides an overview of our real-time lecture translation system, Section 6 concludes this paper.

2. DEVELOPMENT AND EVALUATION DATA

For the automatic speech recognition and statistical machine translation experiments on lectures, we selected three different lectures as development and evaluation data. The three lectures were given in English by the same non-native speaker on different topics. All lectures were recorded with close talking microphones [3].

Dev: A 24min talk that was held to give a broad overview of current research projects in our lab and is therefore ideal as development set.

t035: A 35min talk held as a conference key-note, only partly covered by the Dev talk, which gives us the opportunity to evaluate how our system behaves on an unseen domain.

t036+: A 31min talk on the same topic as t035, but held in a different environmental setting and situation, which allows us to evaluate the robustness of our system.

For the ASR experiments we used the seminar part of the NIST RT-06S development data and the 2006 EPPS development data as additional data sources.

3. SPEECH RECOGNITION

In this section we first compare the 2006 evaluation systems for European Parliament Plenary Sessions [19] and CHIL seminars [4] and describe the development of a single system, which performs almost as good as the evaluation systems on both domains respectively. This is followed by the presentation of the system's performance on the lecture domain. Lectures are an ideal showcase for speaker and domain adaptation tasks, since the lecturer and the topic might be known in advance [1]. Therefore, we describe acoustic and language model adaptation results in the last part of this section. Different from the work [3] we will in this paper take the 2006 EPPS evaluation into consideration for the development of our lecture recognition system.

All speech recognition experiments were done using the Janus Recognition Toolkit (JRTk) featuring the Ibis decoder [17]. For language modeling, we used the SRI Language Modeling Toolkit (SRILM) [18].

3.1. Data

For acoustic model training, we selected the following corpora: ICSI and NIST meeting recordings [9, 12], TED lectures [11], CHIL seminars [25], and European Parliament Plenary Sessions (EPPS) [8]. Because of the results given in [4] we have neither used the ISL meeting corpus nor the Hub4 Broadcast News corpus due to their channel mismatch: both corpora were recorded with lapel microphones. Table 1 gives an overview of the total amount of speech in the different corpora. More information about the respective corpora can be found in the cited literature.

	ICSI	NIST	TED	CHIL	EPPS
speakers	463	77	52	67	1894
duration	72h	13h	13h	10h	80h

Table 1: Number of speakers and total amount of speech data used for acoustic model training.

For language model training, some additional text data was used on top of the 2006 evaluation systems' [4, 19] language model training data. Altogether, the following corpora were available: Talks, text documents from TC-STAR and CHIL, EPPS transcripts, EPPS final text editions, AMI meeting data, non-AMI meeting data (ISL, ICSI, NIST), TED lectures, CHIL seminars, broadcast news data, UN (United Nations) text data released by ELDA, recent proceedings data (2002 - 2005), web data from UWash (related to ISL, ICSI, and NIST meetings) and web data collected for RT-06S (related to CHIL seminars). Table 2 shows the amount of words available for each corpus. More details can be found in [4, 19].

3.2. System Description

The acoustic models used in the experiments below were all trained in the same fashion, resulting in a size of 16,000 distributions over 4,000 models, with a maximum of 64 Gaussians per model. These models were all based on the same quint-phone context decision tree and phoneme set that was used for the RT-06S evaluation system. Furthermore, the acoustic model training setup was taken from the RT-06S system: (1) a first incremental growing of Gaussians, (2) estimation of the global STC transform [7], followed by (3) a second incremental growing of Gaussians. To train the distributions for the semi-continuous system and to compensate for the occasionally worse fixed-state alignments, 2 iterations of Viterbi training were performed (4), and finally, 4 additional iterations of SAT Viterbi training by using constrained MLLR in the feature space (FSA) [6] were applied for the SAT models (5). More details can be found in [4].

Different from [4] and [19] we used a less complicated decoding setup. Instead of doing cross adaptation between systems trained with different phoneme sets and front-ends, we simply used our standard phoneme set and MFCC FFT front-end with a 42-dimensional feature space after linear discriminant analysis (LDA) and a global STC transform with utterance-based cepstral mean subtraction (CMS).

3.3. Baseline Experiments and Comparisons

The goal was to build a single acoustic model for both domains, EPPS and CHIL seminars and to finally use this acoustic model on the lecture data. For this, we compared different acoustic models trained on different subsets of the acoustic training material described in 3.1. All subsets contain the CHIL corpus, which is therefore not explicitly mentioned in the table rows below.

	talks	docs	eppsS	eppsT	nAMI	AMI	TED	CHIL	BN	UN	proc	UWash	wCHIL
words	93k	192k	750k	33M	1.1M	200k	98k	45k	131M	42M	23M	147M	146M
EPPS			35%	54%					9%	2%			
CHIL					15%	8%	0.6%	25%	0.8%		24%	12%	15%
Dev	36%	1%		12%			3%		8%		9%	11%	19%

Table 2: Amount of language model training data in words together with their interpolation weights for the different domains. 'Dev' is the lecture development set as described in Section 2. Empty cells indicate that the data was not useful for that domain.

We used a three pass decoding setup. As in [4], the first pass uses incremental speaker based vocal tract length normalization (VTLN) and constrained MLLR estimation and is decoded with the semi continuous models (4) using tight search beams. The second pass uses the same semi continuous acoustic models as pass one, but before decoding, MLLR [13] adaptation together with an estimation of fixed VTLN and constrained MLLR parameters is performed. For this, the confidence weighted hypotheses of the previous pass are being used. For the third pass, the FSA-SAT acoustic models (5) are used together with the same adaptation scheme applied in pass two. After that, confusion network combination (CNC) [14] is being performed, using the lattices of the third pass only. We used exactly the same decoding dictionaries and language models as for the EPPS and RT-06S evaluation systems.

CHIL Seminars For the CHIL seminars we used the same language models and dictionaries as described in [4]. The language model was trained on AMI and non-AMI meetings, TED, some CHIL data, BN, proceedings and web data related to meetings and CHIL lectures. The interpolation weights, which were tuned on held-out CHIL data are shown in Table 2. The language model has a perplexity of 130 on the RT-06S development data, while 16% 4-grams, 41% 3-grams, 39% 2-grams, and 4% 1-grams were used. The dictionary consists of about 59k pronunciation variants defined over a vocabulary of 52k. It has an out-of-vocabulary (OOV) rate of 0.65 on the RT-06S development data.

As can be seen in table 3 for the above described different system passes, acoustic models trained on EPPS alone or additionally including TED (TED+EPPS) is significant worse than the other two systems. The performance of the two other systems is nearly identical, which means that adding the EPPS data to the acoustic model training data used in RT-06 (ICSI+NIST+TED) does not hurt (but also does not improve the overall results).

CHIL	1st	2nd	3rd	cnc
EPPS	40.3	–.-	–.-	–.-
TED+EPPS	38.7	–.-	–.-	–.-
ICSI+NIST+TED+EPPS	34.1	27.5	26.2	25.5
ICSI+NIST+TED	34.0	27.1	26.0	25.5

Table 3: Results on the RT06 development data. The CHIL data was used in all systems for AM training.

European Parliament Plenary Sessions For the European Parliament Plenary Sessions we used the language models and dictionaries as described in [19]. The language model was trained on EPPS transcriptions and final text editions, BN, and UN and achieved a perplexity of 93 on the 2006 EPPS development data, with 29% 4-grams, 36% 3-grams, 32% 2-grams, and 4% 1-grams. The interpolation weights were tuned on the 2005 EPPS development data and are shown in Table 2. The dictionary for EPPS consists of 45k pronunciations over a vocabulary of 40k and has an OOV-Rate of 0.43 on the 2006 EPPS development data.

As can be seen in Table 4 the system trained without EPPS (ICSI+NIST+TED) performs worst. Furthermore, compared to the acoustic model used for the 2006 EPPS evaluation (MS23), the acoustic model training setup developed for RT-06S is significantly better (MS23 vs. EPPS rows). An additional gain can be seen by adding TED, which is a corpus containing European English as well. By adding the meeting data, the system improves not further, instead it ranks between the EPPS and TED+EPPS systems. Nevertheless, after doing confusion network combination, it gives the same results compared to the TED+EPPS system.

	1st	2nd	3rd	cnc
MS23	22.6	–.-	–.-	–.-
EPPS	20.8	15.4	14.7	14.5
TED+EPPS	20.1	14.8	14.3	14.1
ICSI+NIST+TED+EPPS	20.6	15.1	14.6	14.1
ICSI+NIST+TED	29.1	–.-	–.-	–.-

Table 4: Results on the 2006 EPPS development data. The CHIL data was used in all systems for AM training, except for MS23. MS23 specifies the 2006 EPPS evaluation setup.

Compared to the CHIL seminars, the EPPS results are much better. The reason for that lies in the huge amount of acoustic and language model in-domain training data available for EPPS, while only a very small amount of in-domain data is available for CHIL. Furthermore, the language used in the European Parliament is more formal and therefore less spontaneous. This leads also to a better OOV-rate and language model perplexity with a higher n-gram coverage for larger n.

3.4. Lecture Domain

Based on the perplexities and OOV-rates on Dev shown in Table 5 we selected the language model and dictionary built for the CHIL seminars for our baseline experiments. Not surprisingly, this selection holds also for the evaluation talks. The EPPS language model and vocabulary is, due to the large amount of in-domain data, too specific. The OOV-rates of the RT-06S (CHIL) vocabulary and for t036+ are surprisingly low – the only explanation for that is this talk is not very specific.

	Dev		t035		t036+	
	PPL	OOV	PPL	OOV	PPL	OOV
CHIL	173	0.22	117	0.27	186	0.09
EPPS	205	1.29	230	1.83	229	1.72

Table 5: Perplexities (PPL) and OOV-rates of the CHIL and EPPS language models and vocabularies.

As can be seen in Table 6, the acoustic model trained on all data performs significantly better than the other models. For this reason we selected this model for our further experiments. The baseline results on the lecture evaluation talks are shown in Table 7. With the training setup developed for RT-06S we significantly improved our results compared to the acoustic models developed in [3] (MS11 column in Table 7). Furthermore, it can be seen that the system performs quite well on unseen domains (t035) and different environments (t036+).

Model Adaptation Experiments The baseline experiments were performed with unsupervised adaptation. As mentioned above, for lectures, speaker and topic are often known in advance. Therefore, the lecture domain is ideal for applying supervised acoustic and language model adaptation. As will be shown, this allows us to reduce the decoding setup from three to only one single decoding pass without any loss in performance and is the first step towards a real-time lecture translation system.

For acoustic model adaptation an additional amount of around 7 hours of speech for the same speaker was available. For the adaptation experiments subsets of this data with different lengths were used to compute VTLN and constrained MLLR (FSA) parameters and to perform model based MLLR adaptation. The results can be seen in Table 8. While the adaptation works quite well on the evaluation talks – the 7hrs results are similar to those achieved after CNC with the baseline systems – the results on the

	1st	2nd	3rd	cnc
EPPS	23.9	–.-	–.-	–.-
TED+EPPS	23.4	–.-	–.-	–.-
ICSI+NIST+TED+EPPS	21.4	16.2	15.0	15.5
ICSI+NIST+TED	24.3	–.-	–.-	–.-

Table 6: Baseline results on Dev with the CHIL dictionary and language model. The CHIL data was used in all systems for acoustic model training.

	1st	2nd	3rd	cnc	MS11
t035	17.3	12.6	12.1	12.2	12.7
t036+	16.7	12.0	11.6	11.5	12.4

Table 7: Baseline results on the evaluation talks t035 and t036+. The MS11 column contains the final (CNC) results with the acoustic model trained in [3].

	0.5hrs	1.5hrs	3.5hrs	7hrs	sup
Dev	20.9	20.0	19.5	18.9	12.0
t035	14.2	13.1	12.6	12.1	10.1
t036+	13.3	12.3	11.5	10.7	9.3

Table 8: Acoustic model adaptation results with different amounts of adaptation data. In the column 'sup', supervised adaptation was performed on the particular talk itself.

Dev talk are significantly worse. This is due to a large channel mismatch between the adaptation material and the Dev talk. To confirm this, we adapted on the particular talk itself and achieved reasonable results for all talks (see column 'sup' in Table 8). It can also be seen, that doubling the adaptation data results in a relative gain of about 0.5% in WER.

For language model adaptation we did an initial experiment by tuning the interpolation weights and selecting the different corpora with respect to the lecture domain. The interpolation weights, tuned on some held-out data and the selected corpora can be seen in Table 2. Thereby the perplexity on the Dev talk could only be reduced slightly from 173 to 168. Nevertheless we saw significant gains in WER on all lectures, which are reported in Table 9.

4. STATISTICAL MACHINE TRANSLATION

In this section, we describe the statistical machine translation (SMT) component in our lecture translator that was used to translate the lectures in section 2 from English to Spanish and German. The underlying phrase-based SMT system was originally developed within TC-STAR for translating speeches from the European Parliament Plenary Sessions (EPPS). In these experiments, we used loose coupling, passing the first-best hypothesis from the recognizer to the translation component. Translation results are reported using the well known evaluation metrics BLEU [16] and NIST [15]. All MT scores were calculated using case-insensitive scoring and one reference translation per test set.

	unadapted	adapted	PPL
Dev	18.9	16.1	168
t035	12.1	10.5	165
t036+	10.7	9.1	193

Table 9: Language model adaptation results on top of the acoustic model adaptation on 7hrs of speech. Perplexities should be compared with Table 5.

4.1. Phrase Alignment

To find a translation for a source phrase $\tilde{f} = f_1 ... f_l$ we restrict the general word alignment: Words inside the source phrase align to words inside the target phrase, and words outside the source phrase align to words outside the target phrase. This constrained alignment probability is calculated using the well-known IBM1 word alignment model, but the summation of the target words is restricted to the appropriate regions in the target sentence. Also, the position alignment probabilities are adjusted accordingly [23]. Optimization is over the target side boundaries i_1 and i_2.

$$\begin{aligned} p_{i_1,i_2}(f|e) = & \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \\ & \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{1}{k} p(f_j|e_i) \\ & \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \end{aligned} \quad (1)$$

Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words.

To find the optimal target phrase we interpolate the log probabilities and take the pair (i_1, i_2) that gives the highest probability. The interpolation factor c can be estimated on a development test set.

The scores calculated in the phrase alignment are alignment scores for the entire sentence. As phrase translation probabilities we use the second term in Eqn. 1.

4.2. Decoder

The beam search decoder combines all model scores to find the best translation. In these experiments, the different models used were: (1) The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus according to the new alignment method described in this paper. (2) A trigram language model. The SRI language model toolkit was used to train the models [18]. (3) A word reordering model, which assigns higher costs to longer distance reordering. We use the jump probabilities $p(j|j')$ of the HMM word alignment model [24] where j is the current position in the source sentence and j' is the previous position. (4) Simple word and phrase count models. The former is essentially used to compensate for the tendency of the language model to prefer shorter translations, while the latter can be used to give preference to longer phrases. For each model a scaling factor can be used to modify the contribution of this model to the overall score.

The decoding process is organized into two stages: First, the word-to-word and phrase-to-phrase translations and, if available, other specific information like named entity translation tables are inserted into a translation lattice. In the second step, we find the best combinations of these partial translations, such that every word in the source sentence is covered exactly once. This amounts to doing a best path search through the translation lattice, which is extended to allow for word reordering: Decoding proceeds essentially along the source sentence. At each step, however, the next word or phrase to be translated may be selected from all words laying or phrases starting within a given look-ahead window from the current position [22].

4.3. Training Data

For training the baseline translation systems, the parallel EPPS corpus was used. For English-Spanish, a version was created by RWTH Aachen within TC-STAR [8]. The English-to-German models were trained on the EPPS data as provided by Philipp Koehn [10].

In addition, a small number of lectures similar in style to our development and evaluation data was collected, transcribed, and translated into Spanish and German. Altogether, parallel lecture corpora of about 12,000 words were available in each language.

4.4. Model Adaptation

Adapting the MT component of our EPPS translation system towards the more conversational style of lectures was accomplished by a higher weighting of the available lecture data in two different ways. First, for computing the translation models, the small lecture corpora were multiplied several times and added to the original EPPS training data. This yielded a small increase in MT scores.

Secondly, for (target) language model computation, a small tri-gram LM was computed on t035 and then interpolated with the original EPPS language model, whereas the interpolation weight was chosen in order to minimize the perplexity on the development set. In this manner the perplexity on the Dev talk could be reduced from 645 to 394 for German and from 543 to 403 for Spanish. To further adapt the target language models, we collected Spanish and German web data with the help of tools provided by the University of Washington [21]. A small amount of the used search queries were hand written, however, most search queries were automatically created by using the most frequent tri-grams found in the Dev talk. Approximately $1/4$ of all development set tri-grams were used for this. The German and Spanish web corpora collected in this manner consisted of 175M words and 120M words, respectively. The web corpora were again added to the existing LMs by interpolation, which yielded a perplexity of 200 for German and 134 for Spanish. The corresponding perplexities on the t036+ talks are 617 and 227, respectively.

The effects of translation model and language model adaptation, as well as the results of the final system, combining both adaptation steps, are shown in tables 10 and 11 for English-to-Spanish and English-to-German, respectively.

system	NIST	Bleu
baseline (EPPS)	4.71 (5.61)	15.41 (20.54)
TM-adaptation	4.78 (5.67)	16.05 (21.43)
LM-adaptation	5.10 (5.99)	17.58 (22.90)
final system	5.22 (6.11)	18.57 (24.00)

Table 10: English-Spanish lecture translation system on t036+. Translation results on manual transcripts are shown in brackets.

system	NIST	Bleu
baseline (EPPS)	4.00 (4.71)	09.32 (12.53)
TM-adaptation	4.29 (5.06)	11.01 (14.95)
LM-adaptation	4.37 (5.12)	11.67 (14.96)
final system	4.67 (5.47)	13.22 (17.25)

Table 11: English-German lecture translation system on t036+. Translation results on manual transcripts are shown in brackets.

The significantly lower MT scores for the English-to-German translation direction are mostly due to long distance dependencies and compound words which are inherent to the German language.

In absolute terms, the translation performance on this difficult task is still quite poor when compared with tasks for which large amounts of training data similar in style is available, such as the TC-STAR EPPS task. Nevertheless, small amounts of lecture data were sufficient to significantly improve performance, especially when amplified by using language model adaptation with similar web data.

5. The Real-time Lecture Translation System

For our current version of a real-time lecture translation system, which simultaneously translates lectures given in English into Spanish and German, we integrated the above described speech recognition and machine translation systems together with a sentence segmentation component and a speech synthesis into a client-server framework similar to the one described in [5].

To reach real-time end-to-end performance, we had to tune the above described single pass speech recognizer to run faster than real-time, by further restricting the beam search, which resulted in an increase in WER to about 13% on the evaluation talks. The other system components did not need further tuning.

To keep the latency of the system as short as possible, the speech recognizer already starts to decode, while the speaker is talking and continously returns partial back traces with first best hypotheses. Since the machine translation awaits complete sentences as input, we merged the partial hypotheses together, and resegmented them to sentence like segments. This means, that different from other speech transcription systems no speech segmentation was performed before processing it by the speech recognizer, instead it was done afterwards, to have the ability to tune the segmentation boundaries with respect to optimal machine translation performance. Currently, the segmentation is done at silence regions only, whereby additional thresholds are defined to produce segments with a length of about five to ten words. Thereby, the latency of the system could be limited to a maximum of about five seconds. We plan for more sophisticated segmentation algorithms in the future.

An overview of the real-time lecture translation system is given in Figure 1. As can be seen, the system can deliver the output in different ways:

Subtitles: Simultaneous translations can be projected to the wall as subtitles. This is suitable if the number of output languages is small.

Heads-Up Display Goggles: When there is not enough space on a wall or canvas, heads-up display goggles can be worn to see the simultaneous translation as subtitles. Furthermore, other participants are not disturbed by the subtitles.

Targeted Audio Another solution for providing the simultaneous translation without disturbing others is the so-called targeted audio device [2]. The targeted audio device is a beam-steered loudspeaker, consisting of several small ultrasound loudspeakers. It outputs audio in a beam with a width of about 1-2 meters. People sitting within the beam are able to hear everything, people outside the beam do not. In future applications, several such targeted audio devices could be assigned in various languages to accommodate each participant in the lecture room.

6. Conclusion

In this paper, we presented our work in taking first steps towards building open domain speech translation systems. We have successfully developed an ASR system for lectures by merging the evaluation systems for European Parliament Plenary Sessions and CHIL seminars. Furthermore, we combined the resulting system with the translation system used in TC-STAR to translate lectures on a new domain from English to Spanish and German.

The ASR system performance exceeds our expectations, demonstrating the feasibility of designing open domain recognition systems. For translation, lectures still pose a significant challenge. Nevertheless, small amounts of lecture data were sufficient to significantly improve performance, especially when amplified by using language model adaptation with similar web data.

7. Acknowledgments

This work has been funded by the *European Union* under the integrated projects CHIL – Computers in the Human Interaction Loop – (Grant number IST-506909) and TC-

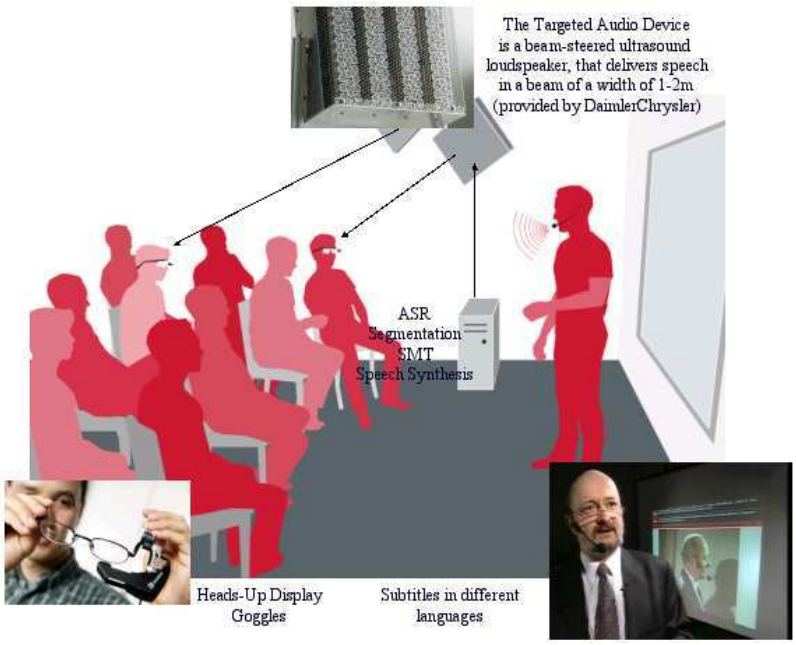


Figure 1: The lecture transcription system.

STAR – Technology and Corpora for Speech to Speech Translation – (Grant number IST-506738). The authors would like to thank Susanne Burger, Maria Kernecker, Tim Notari, and Silja Hildebrand for transcribing and translating the lecture development and evaluation data and Florian Kraft for providing the language model data used for the RT-06S evaluation system.

BIBLIOGRAPHIE

[1] M. Cettolo, F. Brugnara, and M. Federico. Advances in the Automatic Transcription of Lectures. In *ICASSP*, Montreal, Canada, 2004.

[2] K. Linhard D. Olszewski, F. Prasetyo. Steerable Highly Directional Audio Beam Louspeaker. In *Proc. of the Interspeech*, Lisboa, Portugal, September 2006.

[3] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel. Open Domain Speech Recognition & Translation: Lectures and Speeches. In *ICASSP*, Toulouse, France, 2006.

[4] C. Fügen, M. Wölfel, J. W. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani. Advances in Lecture Recognition: The ISL RT-06S Evaluation System. In *submitted to Interspeech 2006*, Pittsburgh, PA, USA, September 2006.

[5] Christian Fügen, Martin Westphal, Mike Schneider, Tanja Schultz, and Alex Waibel. LingWear: A Mobile Tourist Information System. In *Proc. of the Human Language Technology Conf. (HLT)*, San Diego, California, March 2001. NIST.

[6] M. J. F. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical report, Cambridge University, Cambridge, United Kingdom, 1997.

[7] M. J. F. Gales. Semi-tied covariance matrices. In *ICASSP*, 1998.

[8] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney. Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus. *ICASSP*, 2005.

[9] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. The ICSI Meeting Project: Ressources and Research. In *Proc. of the ICASSP Meeting Recognition Workshop*, Montreal, Canada, May 2004. NIST.

[10] P. Koehn. Europarl: A Multilingual Corpus for Evaluation of Machine Translation, 2003. http://people.csail.mit.edu/koehn/publications/europarl.

[11] L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann. The Translanguage English Database TED. In *ICSLP*, volume LDC2002S04, Yokohama, September 1994. LDC.

[12] Linguistic Data Consortium (LDC). ICSI, ISL and NIST Meeting Speech Coprora at LDC, 2004. http://www.ldc.upenn.edu catalog IDs LDC2004S02, LDC2004S05, LDC2004S09.

[13] C. J. Leggetter and P. C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185, 1995.

[14] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus among Words: Lattice-based Word Error Minimization. In *EUROSPEECH*, 1999.

[15] NIST. NIST MT evaluation kit version 11a, 2004. http://www.nist.gov/speech/tests/mt.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center, 2002.

[17] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment. In *ASRU*, Trento, Italy, 2001.

[18] A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *ICSLP*, Denver, Colorado, USA, 2002.

[19] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, F. Kraft Q. Jin, M. Paulik, M. Raab, Y.-C. Tam, and M. Wölfel. The ISL TC-STAR Spring 2006 ASR Evaluation Systems. In *submitted to the TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, June 2006.

[20] TC-Star. Technology and corpora for speech to speech translation, 2004. http://www.tc-star.org.

[21] UWash. University of Washington, web data collection scripts, 2006. http://ssli.ee.washington.edu/projects/ears/WebData/web_data_collection.html.

[22] S. Vogel. SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.

[23] S. Vogel. PESA: Phrase Pair Extraction as Sentence Splitting. In *Machine Translation Summit 2005*, Thailand, 2005.

[24] S. Vogel, H. Ney, and C. Tillmann. HMM-based Word Alignment in Statistical Translation. In *COLING 96*, pages 836–841, Copenhagen, 1996.

[25] A. Waibel, H. Steusloff, and R. Stiefelhagen. CHIL – Computers in the Human Interaction Loop. In *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, April 2004. http://chil.server.de.

Session XI

Compréhension automatique

Mercredi 14 juin 2006 - 10h00 11h00

Mesure de confiance de relation sémantique dans le cadre d'un modèle de langage sémantique

Catherine Kobus, Géraldine Damnati et Lionel Delphin-Poulat

France Télécom R&D
2, avenue Pierre Marzin 22307 Lannion - France

ABSTRACT

This article proposes a new confidence measure estimated for concept hypotheses given by a semantic language model. This confidence measure is based upon the ontology and the semantic relations linking concepts of a dialog application. It aims at measuring how high a concept hypothesis is related to the other hypotheses of an utterance. The semantic relation confidence measure is evaluated alone and in combination with a classical acoustic confidence measure. It is shown that the two confidence measures are complementary and yield good performance in terms of cross entropy relative reduction.

1. INTRODUCTION

Les mesures de confiance sont souvent utilisées en reconnaissance de parole afin de traduire la fiabilité de la réponse (ou hypothèse) fournie par un système de reconnaissance vocale. Les applications des mesures de confiance sont multiples ; elles peuvent être utilisées dans le cadre d'une stratégie de rejet des hypothèses peu fiables ou d'une adaptation du dialogue (demande de confirmation ou non de la phrase prononcée par l'utilisateur), etc. Les mesures de confiance peuvent être évaluées à plusieurs niveaux suivant le but recherché : au niveau de l'hypothèse du mot, du concept ou de la phrase. Il est crucial, dans les applications de reconnaissance vocale et en particulier les applications de dialogue, de déterminer le degré de fiabilité des concepts reconnus.
Les mesures de confiance présentées dans cet article sont calculées sur des hypothèses de concepts, issues d'un modèle de langage sémantique [4], appliqué à un graphe de mots en seconde passe de reconnaissance, dans le cadre d'une application de dialogue. L'application utilisée est liée au domaine bancaire ; elle permet à un utilisateur de consulter ses comptes, ses ordres effectués ou d'effectuer des transactions boursières. Plusieurs méthodes d'estimation de mesures de confiance au niveau des concepts ont déjà été proposées. Dans [3], une mesure basée sur la probabilité *a posteriori* d'un concept est estimée à partir d'un graphe de concepts. [5] introduit une méthode basée sur le consensus de plusieurs classifieurs sémantiques pour valider une hypothèse conceptuelle. [6] incorpore de l'information d'ordre sémantique dans le calcul d'une mesure de confiance sur les concepts, basée sur la probabilité *a posteriori*. La mesure de confiance de relation sémantique présentée dans cet article utilise l'ontologie de l'application de dialogue et, plus précisément, les différentes relations sémantiques pouvant lier les concepts de l'application, ce qui permet de mesurer un indice de cohérence pour un concept d'une phrase par rapport aux autres concepts de cette même phrase. L'article s'organise de la façon suivante : la première partie définit la mesure de confiance de relation sémantique, et rappelle la définition de la mesure de confiance acoustique. La partie suivante présente la régression logistique utilisée pour calibrer et combiner les mesures de confiance. Enfin sont présentés le contexte expérimental de l'application de dialogue, ainsi que les différents résultats obtenus avec les mesures de confiance (utilisées individuellement ou en combinaison).

2. MESURES DE CONFIANCE

2.1. Mesure de confiance de relation sémantique

L'idée est d'utiliser la source de connaissance que représentent les relations possibles pouvant exister entre les différents concepts de l'application, afin de vérifier qu'une hypothèse de concept, proposée par le modèle de langage sémantique, est cohérente avec l'ensemble des hypothèses de concepts de la phrase.

Ontologie de l'application de dialogue Une ontologie est un ensemble structuré de concepts. Elle décrit comment les différents concepts de l'application sont organisés et liés par des relations sémantiques. L'ontologie pour l'application utilisée est décrite avec le langage KL-One, langage de représentation des connaissances, introduit par Brachman [1]. Ce langage correspond à la première implémentation des réseaux structurés d'héritage ou SIN (pour *Structured Inheritance Network*). Le langage KL-One distingue d'une part les classes conceptuelles, d'autre part les concepts, instances des classes conceptuelles (chaque concept de l'application est une instance d'une seule classe). Il décrit également les relations sémantiques pouvant exister entre certaines de ces classes. Les relations sémantiques sont des relations binaires liant deux classes conceptuelles. Les concepts, instances des classes conceptuelles en relation, sont liés par la même relation. De plus, une relation sémantique liant deux concepts de l'application ne tient pas compte de l'ordre dans lequel ils apparaissent.
Une relation, liant deux classes génériques, est définie de façon unique par la donnée de ces deux classes. Par convention, toute classe conceptuelle est en relation avec elle-même.
La figure 1 représente des exemples de classes conceptuelles en relation dans l'ontologie ; les classes conceptuelles $claNiveau$ et $claIndice$ sont liées par la relation $R_{claNiveau\_claIndice}$, ainsi que les concepts, instances de chacune de ces classes. Dans le cadre de nos travaux effec-

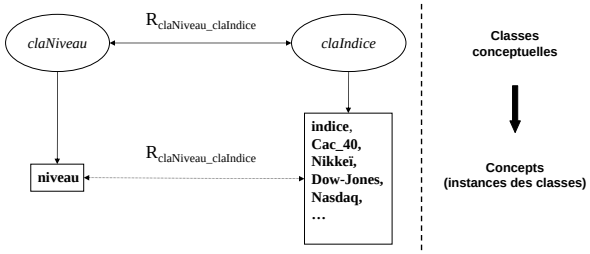


FIG. 1: Exemple de relation décrite par KL-ONE.

tués sur les mesures de confiance, un type de relation particulier a été introduit ; il s'agit de la relation dite "nulle" associée à chaque classe conceptuelle. Un concept, instance d'une classe conceptuelle, est dit en relation nulle lorsqu'il apparaît seul dans un énoncé. Les relations nulles peuvent être vues comme des relations liant chacune des classes conceptuelles à la classe conceptuelle "vide".

Modèle probabiliste On veut utiliser les différentes relations sémantiques de l'ontologie afin de définir une distribution de probabilité P^{Rel}, telle que, pour deux concepts c_i et c_j de l'application, $P^{Rel}(c_i, c_j) = 0$ si les concepts ne sont liés par aucune relation sémantique, et que $P^{Rel}(c_i, c_j)$ soit fonction de la probabilité de la relation sémantique les reliant, dans le cas contraire. Nous introduisons une fonction γ, qui pour chaque concept c_i, donne l'unique classe conceptuelle dont il est instance.
Chaque concept c_i étant instance d'une unique classe conceptuelle $\gamma(c_i)$, la probabilité de relation $P^{Rel}(c_i, c_j)$ est définie comme suit :

$$\begin{aligned} P^{Rel}(c_i, c_j) &= P(c_i, c_j|\gamma(c_i), \gamma(c_j)) \cdot P(\gamma(c_i), \gamma(c_j)) \\ &\simeq P(c_i|\gamma(c_i)) \cdot P(c_j|\gamma(c_j)) \\ &\quad \cdot P(\gamma(c_i), \gamma(c_j)) \end{aligned} \tag{1}$$

Nous introduisons alors la relation sémantique de l'ontologie, liant les classes conceptuelles $\gamma(c_i)$ et $\gamma(c_j)$ et notée $R_{\gamma(c_i),\gamma(c_j)}$. La relation sémantique ne tient pas compte de l'ordre d'apparition des concepts dans la phrase ; on pose ainsi :

$$P(\gamma(c_i), \gamma(c_j)) = P(\gamma(c_j), \gamma(c_i)) = \frac{1}{2} P(R_{\gamma(c_i),\gamma(c_j)}) \tag{2}$$

La probabilité de relation $P^{Rel}(c_i, c_j)$ devient alors :

$$P^{Rel}(c_i, c_j) = \frac{1}{2} P(c_i|\gamma(c_i)) P(c_j|\gamma(c_j)) P(R_{\gamma(c_i),\gamma(c_j)}) \tag{3}$$

On peut ainsi définir, pour chaque concept c_i de l'application, une probabilité $P^{Rel}(c_i)$, probabilité que le concept c_i soit en relation avec les autres concepts de l'application. Par définition,

$$P^{Rel}(c_i) = \sum_{c_j} P^{Rel}(c_i, c_j) \tag{4}$$

Les différentes probabilités de relation et d'instanciation des concepts sont estimées par comptage puis lissées à partir d'un corpus d'apprentissage étiqueté en concepts.

Information mutuelle Pour définir une mesure de confiance, nous calculons la quantité d'information mutuelle, notée $i^{Rel}(c_i; c_j)$, apportée par la co-occurrence des concepts c_i et c_j en relation selon la distribution de probabilité P^{Rel}.

$$i^{Rel}(c_i; c_j) = \log \frac{P^{Rel}(c_i, c_j)}{P^{Rel}(c_i) \cdot P^{Rel}(c_j)} \tag{5}$$

Cas d'une hypothèse de phrase contenant au moins deux concepts Soit W une hypothèse de phrase, déduite par le modèle de langage sémantique [4] et contenant n concepts ($n > 2$) $c_1, ...c_n$. Soient c_i et c_j deux concepts de cette hypothèse, liés par la relation R dans l'ontologie. L'information mutuelle $i^{Rel}(c_i; c_j)$ permet de mesurer l'influence de l'occurrence du concept c_i sur l'occurrence du concept c_j dans la même phrase (et inversement). La mesure de confiance de relation sémantique est calculée, pour chaque hypothèse de concept c_i, en faisant la moyenne de l'ensemble des informations mutuelles de relation du concept c_i avec les autres concepts de l'hypothèse. L'information mutuelle étant nulle pour deux concepts qui ne sont pas en relation, la moyenne sur l'ensemble des relations possibles *a priori* dans la phrase permet de tenir compte, dans la mesure, des concepts non liés sémantiquement. La mesure de confiance de relation pour un concept c_i d'une hypothèse constituée de n concepts ($n > 2$), notée $MC^{Rel}(c_i)$, est définie de la façon suivante :

$$MC^{Rel}(c_i) = \frac{1}{n-1} \sum_{j \neq i / Rel(c_i,c_j)=1}^{n} i^{Rel}(c_i; c_j) \tag{6}$$

La notation $Rel(c_i, c_j) = 1$ signifie que les concepts c_i et c_j sont en relation dans l'ontologie.

Cas d'une hypothèse constituée d'un seul concept L'ajout de la relation "nulle" pour chaque classe générique conceptuelle (effective lorsqu'un concept apparaît seul comme hypothèse du modèle de langage sémantique) permet d'étendre la définition de la mesure de confiance au cas d'un concept isolé. L'apprentissage des probabilités sur les relations de l'ontologie permet entre autres de constater les fréquences d'observation de certains concepts suivant les relations. Notamment, certains concepts sont plus fréquemment observés en tant qu'hypothèse unique du modèle de langage sémantique que d'autres. C'est pourquoi une mesure de confiance de relation peut être intéressante dans le cas d'hypothèses de concepts isolés.
Une relation "nulle", notée $R_\epsilon(X)$, peut être vue comme une relation liant une classe générique conceptuelle X à une classe "nulle", noté C_ϵ. Le nombre de relations possible étant de 1 pour des hypothèses de concepts isolés, la mesure de confiance de relation $MC^{Rel}(c)$ associé à un concept c isolé est égale à l'information mutuelle apportée par l'occurrence d'un concept c en relation "nulle" :

$$MC^{Rel}(c) = i^{Rel}(c; c_\epsilon) = \log \frac{P^{Rel}(c, c_\epsilon)}{P^{Rel}(c) \cdot P^{Rel}(c_\epsilon)} \tag{7}$$

Dans l'équation 7, la probabilité $P^{Rel}(C_\epsilon)$ que le concept "nul" soit en relation, est en fait la probabilité d'occurrence de la relation "nulle", égale à la fréquence d'occurrences d'hypothèses de concepts isolés sur le corpus d'apprentissage. La probabilité $P^{Rel}(c, c_\epsilon)$ peut s'approximer de la façon suivante :

$$P^{Rel}(c, c_\epsilon) \approx P(c|\gamma(c)) \cdot P(R_\epsilon(\gamma(c))) \tag{8}$$

La probabilité $P(R_\epsilon(\gamma(c)))$ représente la probabilité de la relation "nulle" pour la classe $\gamma(c)$.

2.2. Mesure de confiance acoustique

Une mesure de confiance acoustique est également utilisée. Cette mesure compare le score de vraisemblance fourni par le modèle acoustique du système de reconnaissance avec le score, qui serait obtenu en utilisant un modèle acoustique simpliste, constitué d'une boucle de phonèmes sans contrainte. Le modèle acoustique du système de reconnaissance vocale est noté $\mathcal{A}_G$, tandis que le modèle constitué d'une boucle de phonèmes est noté $\mathcal{A}_{boucle}$. Soit w un mot hypothèse du système de reconnaissance vocale, reconnu avec le modèle acoustique $\mathcal{A}_G$ et occupant $T(w)$ trames. Le score de vraisemblance associé au signal acoustique X_w correspondant à l'hypothèse de mot w et obtenu avec le modèle acoustique est comparé au score de vraisemblance obtenu pour cette même portion de signal avec le modèle acoustique $\mathcal{A}_{boucle}$, constitué d'une boucle de phonèmes sans contrainte. La différence des vraisemblances obtenues est normalisée par le nombre de trames occupées par le mot ; la mesure de confiance acoustique, associée à l'hypothèse de mot w, notée $MC^{Acous}(w)$, est définie comme suit :

$$MC^{Acous}(w) = \frac{[\log(P(X_w|\mathcal{A}_G) - \log(P(X_w|\mathcal{A}_{boucle}))]}{T(w)} \quad (9)$$

La définition de la mesure de confiance est étendue à une hypothèse de concept([5]). Elle se déduit des mesures de confiance acoustique associées à chaque mot composant le concept. Soit c une hypothèse de concept, constitué de D mots $w_1, w_2...w_D$; la mesure de confiance acoustique associée à l'hypothèse de concept c, notée $MC^{Acous}(c)$, se calcule de la façon suivante :

$$MC^{Acous}(c) = \frac{\sum_{d=1}^{D} T(w_d) \cdot MC^{Acous}(w_d)}{\sum_{d=1}^{D} T(w_d)} \quad (10)$$

3. UTILISATION DES MESURES DE CONFIANCE

Les mesures de confiance définies précédemment sont calculées pour les différentes hypothèses de concepts issues du modèle de langage sémantique ; le jeu de mesures de confiance associé à chaque hypothèse doit permettre d'en estimer la fiabilité, à savoir la probabilité qu'une hypothèse de concept soit correcte étant donné le jeu de mesures de confiance associé.

3.1. Critère d'évaluation

Plusieurs méthodes peuvent être utilisées afin d'évaluer des mesures de confiance ; une méthode possible consiste à mesurer la diminution relative de l'entropie croisée d'un corpus de test engendrée par la mesure de confiance. Soit $\mathcal{C}$ un corpus de test constitué de N hypothèses de concept C_i ; l'entropie croisée de ce corpus, notée H est définie de la façon suivante :

$$H = -\frac{1}{N}(\sum_{i=1}^{N} \delta(C_i) \log p_i + (1 - \delta(C_i)) \log(1 - p_i)) \quad (11)$$

$\delta(c_i)$ est un indicateur égal à 1 si l'hypothèse de concept c_i est correcte, égal à 0 sinon. La probabilité p_i représente la probabilité *a posteriori* que l'hypothèse de concept c_i soit correcte. Sans l'utilisation de mesure de confiance, cette probabilité *a posteriori* est égale à la précision sur la reconnaissance des concepts, notée *Prec.*, sur le corpus de test. $Prec. = \frac{\sum_i \delta(c_i)}{N}$. On définit ainsi, pour un corpus de test donné, une entropie croisée initiale, notée H_{init}.

$$H_{init} = -\textit{Prec.}\log(\textit{Prec.}) - (1 - \textit{Prec.})\log(1 - \textit{Prec.}) \quad (12)$$

En utilisant un ensemble de mesure de confiance $MC = mc_1...mc_n$ ($n \geq 1$), la probabilité *a posteriori* p_i devient $p_i = P(Cor|MC(c_i))$, la probabilité $P(Cor|MC(c_i))$ étant la probabilité qu'une hypothèse de concept c_i soit correcte étant donné un jeu de mesures de confiance associé $MC(c_i)$. Une nouvelle entropie croisée du corpus de test, évaluée avec le jeu de mesures de confiance MC, notée H_{MC}, est alors définie.

$$H_{MC} = -\frac{1}{N}(\sum_{i=1}^{N} \delta(c_i) \log(P(Cor|MC(c_i)) + (1 - \delta(c_i)) \log(1 - P(Cor|MC(c_i)))) \quad (13)$$

Les mesures de confiance sont ainsi évaluées par la diminution relative induite sur l'entropie croisée d'un corpus de test donné, notée ΔH ($\Delta H = \frac{H_{init} - H_{MC}}{H_{init}}$).

3.2. Calibration des mesures de confiance

La méthode d'évaluation choisie requiert la connaissance de la probabilité *a posteriori* qu'une hypothèse de concept soit correcte étant donné le jeu de mesures de confiance associé, à savoir la probabilité $P(Cor|MC(c_i)$. Dans le cas de l'utilisation d'une seule mesure de confiance mc, la probabilité à estimer est la probabilité $P(Cor|mc)$; elle est calibrée au préalable sur un corpus de développement et approximée par régression logistique :

$$P(Cor|mc_1) = \frac{1}{1 + e^{-(a_0 + a_1 \cdot mc_1)}} \quad (14)$$

Les paramètres a_0 et a_1 sont estimés de façon à minimiser l'entropie croisée sur un corpus de développement.

3.3. Combinaison de mesures de confiance

Plusieurs méthodes peuvent être utilisées afin de combiner des mesures de confiance [2], utilisant des sources de connaissance différentes. Dans le cadre de nos travaux, la régression logistique est utilisée afin de combiner les mesures de confiance acoustique et de relation sémantique. La probabilité *a posteriori* qu'une hypothèse de concept soit correcte étant donné les valeurs des mesures de confiance acoustique et de relation sémantique, notées respectivement mc_1 et mc_2 :

$$P(Cor|mc_1, mc_2) = \frac{1}{1 + e^{-(a_0 + a_1 \cdot mc_1 + a_2 \cdot mc_2)}} \quad (15)$$

Les paramètres a_0, a_1 et a_2 sont estimés de façon à minimiser l'entropie croisée du corpus de développement.

4. EVALUATIONS

4.1. Description des données

Les travaux sur le modèle de langage sémantique et sur les mesures de confiance ont été réalisés dans le cadre d'une application de dialogue [4] dans le domaine bancaire permettant à un utilisateur de consulter ses comptes, d'effectuer diverses transactions ou d'obtenir des informations liées à la bourse. Les différentes probabilités de relation et d'instanciation des concepts, utiles au calcul de la mesure de confiance de relation sémantique, ont été estimées (par

TAB. 1: Description des différents corpus de test

	Nb. Conc.	Préc.
Dev_1	1516	12.13
Dev_2+	2956	10.12
$Test\_1$	966	10.15
$Test\_2+$	1894	9.29

comptage puis lissage) sur un corpus d'apprentissage de 24604 énoncés, contenant 34623 concepts. Le modèle de langage sémantique [4], appliqué en seconde passe de reconnaissance à des corpus de test constitués de graphes de mots, donne pour chacun de ces ensembles des hypothèses de concepts. Le premier ensemble, noté Dev, constitué de 4472 hypothèses de concepts, est utilisé comme corpus de développement pour la calibration et la combinaison des mesures de confiance. Le deuxième corpus, noté $Test$, est constitué de 2860 hypothèses de concepts et a été utilisé comme corpus de test pour l'évaluation des mesures de confiance. La mesure de confiance de relation sémantique étant définie de façon différente selon le type de concept (isolé ou non dans la phrase), la question se posait alors de savoir s'il fallait fractionner le corpus de développement en conséquence et calibrer différemment les mesures de confiance pour les concepts isolés et non isolés. Les résultats ont montré que le corpus de développement choisi (fractionné ou non) influençait les performances de la mesure de confiance de relation sémantique et qu'il était préférable de fractionner le corpus de développement lors de la calibration de cette mesure de confiance. Les mesures de confiance ont donc été calibrées différemment pour les concepts isolés et les concepts non isolés. Par la suite, les résultats sont présentés en distinguant les hypothèses de phrases des corpus de test contenant une seule hypothèse de celles contenant au moins deux concepts. Ainsi, Dev_1 (respectivement $Test\_1$) représente l'ensemble des hypothèses du corpus de développement Dev (respectivement du corpus de test $Test$) ne contenant qu'un seul concept, tandis que Dev_2+ (respectivement $Test\_2+$) représente l'ensemble des hypothèses du corpus de développement Dev (respectivement du corpus de test $Test$) en contenant au moins deux. Le tableau 1 donne pour chaque corpus (de développement et de test) le nombre de concepts et la précision sur la reconnaissance des concepts (et de leur valeur).

4.2. Résultats

Les résultats obtenus en termes de diminution relation relative d'entropie croisée sur les différents corpus de test et avec les différentes mesures de confiance (utilisées seules, MC^{Rel} et MC^{Acous} ou en combinaison, $Reg.Log$) sont résumés dans le tableau 2. Les résultats montrent que la mesure de confiance de relation sémantique s'avère être aussi efficace pour les hypothèses de concepts isolés que pour les non-isolés, ce qui conforte le choix de l'introduction de relations dites "nulles" dans notre modèle. La mesure de confiance acoustique reste plus performante, sur les différents corpus, que la mesure de confiance de relation sémantique. Elle semble toutefois être moins performante sur les hypothèses de concepts non isolées (à savoir sur les corpus Dev_2+ et $Test\_2+$). Une analyse détaillée montre que ces corpus contiennent beaucoup plus de concepts courts en termes de nombre de trames (en moyenne 4 fois plus), par rapport aux corpus de concepts isolés (les corpus Dev_1 et $Test\_1$). Une bonne partie de ces concepts courts sont des quantités, pour lesquelles beaucoup d'ambiguïtés peuvent subsister (*"dix"*, *"six"*, etc). L'analyse acoustique est d'autant plus délicate sur des séquences de signal courtes. C'est pourquoi la mesure de confiance acoustique s'avère moins efficace sur ces corpus.

TAB. 2: Evaluations des mesures de confiance en termes de réduction relative d'entropie croisée ($\Delta H(\%)$)

	Dev_1	Dev_2+	$Test\_1$	$Test\_2+$
MC^{Rel}	4.06	4.86	2.37	2.99
MC^{Acous}	13.11	7.00	10.67	7.43
$Reg.Log$	17.00	11.58	13.00	10.27

Les résultats obtenus avec la régression logistique en montrent l'efficacité à combiner des mesures de confiance utilisant des sources de connaissance différentes. Les améliorations apportées individuellement par chaque mesure de confiance sont quasiment cumulées grâce à la régression logistique, ce qui prouve la complémentarité des mesures de confiance.

5. CONCLUSION

Cet article propose une nouvelle mesure de confiance de relation sémantique, calculée pour des hypothèses de concept issues d'un modèle de langage sémantique [4]. Cette mesure utilise l'ontologie d'une application de dialogue et les différentes relations sémantiques liant les concepts de cette dernière, afin de mesurer pour chaque hypothèse de concept un degré de cohérence avec les autres hypothèses de concepts de la même phrase. Une mesure est même introduite pour les concepts isolés, seule hypothèse fournie par le modèle de langage sémantique pour une phrase donnée. Les résultats montrent l'intérêt de cette mesure de confiance, les performances étant d'autant plus importantes en combinant, par régression logistique, cette dernière avec une mesure de confiance acoustique.

RÉFÉRENCES

[1] T. Brachman. What's in a concept : Structural foundations for semantic networks. *International Journal of Man-Machine Studies*, 9 :127–152, 1977.

[2] D. Charlet, G. Mercier, and D. Jouvet. On combining confidence measures for improved rejection of incorrect data. In *Eurospeech*, 2001.

[3] K. Hacioglu and W. Ward. A concept graph based confidence measure. In *ICASSP*, 2002.

[4] C. Kobus, G. Damnati, L. Delphin-Poulat, and R. De Mori. Conceptual language model design for spoken language understanding. In *Eurospeech*, 2005.

[5] C. Raymond, F. Béchet, N. Camelin, R. De Mori, and G. Damnati. Semantic interpretation with error correction. In *ICASSP*, 2005.

[6] R. Sarikaya, Y. Gao, M. Picheny, and H. Erdogan. Semantic confidence measurement for spoken dialog systems. *IEEE Transactions on Speech and Audio Processing*, 2005.

Décodage conceptuel à partir de graphes de mots sur le corpus de dialogue Homme-Machine MEDIA

Christophe Servan, Christian Raymond, Frédéric Béchet, Pascal Nocéra

LIA - Université d'Avignon, BP1228 84911 Avignon cedex 09 France
{christophe.servan,christian.raymond,frederic.bechet,pascal.nocera}@univ-avignon.fr

ABSTRACT

Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. The experiments carried on on the MEDIA corpus show that the performance reached by our approach is better than the traditional sequential approach that looks first for the best sequence of words before looking for the best sequence of concepts.

1. INTRODUCTION

Dans le cadre des applications de dialogue homme-machine, la campagne MEDIA [1] s'est focalisée sur l'évaluation de systèmes de décodage conceptuel permettant d'associer à une séquence de mots une séquence de concepts relatifs au type de dialogue visé. Cette évaluation a été faite sur des transcriptions manuelles de messages audio obtenus grâce à un protocole de type *Magicien d'Oz* sur une tâche de réservation hôtelière. En complément de la campagne MEDIA, cette étude présente les premiers travaux effectués sur le corpus audio MEDIA. En utilisant d'une part le système de Reconnaissance Automatique de la Parole (RAP) SPEERAL [3] et d'autre part le module d'interprétation sémantique développé au LIA [4] nous montrons comment une approche de décodage *intégrée* cherchant directement la meilleure séquence de concepts à partir d'un graphe de mots issus du module de RAP surpasse l'approche séquentielle traditionnelle consistant à détecter les concepts uniquement dans la meilleure hypothèse de phrase sortie par le module de RAP.

Cet article est structuré comme suit : le paragraphe 2 introduit rapidement le corpus MEDIA ; les modèles de RAP développés sur ce corpus sont présentés dans le paragraphe 3 ; le paragraphe 4 rappelle brièvement l'approche du LIA concernant l'interprétation sémantique de message audio ; enfin la partie 5 décrit les résultats des expérimentations effectuées en essayant de répondre aux deux questions suivantes :

– quel est l'impact du taux d'erreurs mots sur les performances du module de décodage conceptuel ?
– quelles sont les différences de performance constatées entre d'une part l'approche de décodage *intégré* proposée dans cette étude et d'autre part l'approche séquentielle traditionnelle.

2. LE CORPUS MEDIA

La campagne d'évaluation MEDIA [1] (programme Technolangue/Evalda) se place dans le cadre de la simulation d'un système d'accès à des informations touristiques et des réservations d'hôtel. Un corpus de 1250 dialogues a été enregistré par ELDA selon un protocole de type *Magicien d'Oz* : 250 locuteurs ont effectué chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. Ce corpus a ensuite été transcrit manuellement, puis annoté sémantiquement selon un dictionnaire sémantique de concepts mis au point par les partenaires du projet. Le dictionnaire sémantique utilisé pour annoter le corpus MEDIA permet d'associer 3 types d'information à un mot ou un groupe de mots :

– tout d'abord une paire attribut-valeur, correspondant à une représentation sémantique *à plat* d'un énoncé ;
– puis un spécifieur qui permet de définir des relations entre les attributs et qui par conséquent peut être utilisé pour construire une représentation hiérarchique de l'interprétation d'un énoncé ;
– enfin une information sur le *mode* attaché à un concept (positif, affirmatif, interrogatif ou optionnel).

n	W^{c_n}	c_n	*valeur*
0	euh	null	
1	oui	réponse	oui
2	l'	LienRef-coRef	singulier
3	hôtel	BDObjet	hotel
4	dont	null	
5	le prix	objet	paiement-montant
6	ne dépasse pas	comparatif-paiement	inferieur
7	cent dix	paiement-montant-ent	110
8	euros	paiement-devise	euro

TAB. 1: Exemple de message annoté du corpus MEDIA

La table 1 présente un exemple de message annoté du corpus. La première colonne correspond au numéro du segment dans le message, la deuxième colonne à la chaîne de mots W^{c_n} porteuse du concept c_n contenu dans la troisième colonne. La quatrième colonne contient la valeur du concept c_n dans la chaîne W^{c_n}. Le dictionnaire sémantique MEDIA contient 83 attributs, auxquels peuvent s'ajouter 19 spécifieurs de relations entre attributs. Le corpus collecté a été découpé en plusieurs lots. Nous utilisons dans cette étude les 4 premiers lots comme corpus d'apprentissage, soit 720 dialogues contenant environ 12K messages, et le lot *Test à blanc* comme corpus de tests contenant 79 dialogues avec 1.3K messages.

3. DÉVELOPPEMENT D'UN SYSTÈME DE RAP SUR LE CORPUS MEDIA

3.1. Apprentissage des modèles

Le décodeur SPEERAL [3] a été utilisé pour transcrire les messages du corpus MEDIA. Ces messages sont enregistrés dans des conditions identiques à celles que l'on peut trouver dans un système mis en service. Les utilisateurs ont effectué leurs appels depuis leur téléphone, fixe ou cellulaire, et la qualité des enregistrements est variable. Les modèles acoustiques téléphoniques utilisés sont ceux développés lors de la campagne d'évaluation ESTER sur la transcription de données radiophoniques, ils ont ensuite été adaptés sur les 720 dialogues des lots 1,2,3,4 du corpus MEDIA par une adaptation de type MAP. Ce sont des modèles triphones.

Le modèle de langage, de type 3-grammes de mots, a été appris sur un corpus extrait des transcriptions manuelles des lots 1,2,3,4. Ce corpus contient un ensemble de 226K mots. Un lexique de 2028 mots a été défini sur ce corpus, il a été phonétisé avec l'outil LIA_PHON[1]. Sur le corpus de test utilisé (lot *Test à blanc* du corpus MEDIA), le taux de mots hors-vocabulaire du lexique choisi est de 1,6%. La perplexité est de 26,5.

Le taux d'erreur mot (ou *Word Error Rate WER*) de la transcription automatique du lot *Test à blanc* avec les modèles présentés (combinaison des modèles acoustiques et linguistiques dans le décodeur SPEERAL) est de 32,2%.

3.2. Graphes de mots

L'approche intégrée de décodage conceptuelle défendue dans cette étude nécessite le traitement de graphes de mots issus du module de RAP. Ces graphes sont produits par le décodeur SPEERAL, toutes les opérations sur les graphes sont ensuite effectuées avec l'ensemble d'outils de manipulation d'automates *AT&T FSM/GRM Library* [2]. Ces graphes nous permettent également de faire varier le WER de la meilleure hypothèse produite par le module de RAP. En effet, un but de cette étude est d'étudier la corrélation entre le taux d'erreur sur les mots et celui sur les concepts. Il est donc intéressant de produire des sorties multiples. Ces sorties sont obtenues par la méthode suivante :

- tout d'abord des graphes de mots sont générés par SPEERAL sur le corpus de test avec les modèles présentés précédemment ; en prenant la meilleure séquence de mots dans ces graphes nous obtenons les hypothèses *baseline* avec un WER moyen de 32,2% ;
- un nouveau modèle de langage (toujours de type 3-grammes) est alors appris, cette fois sur le corpus de test ;
- ce modèle est appliqué aux graphes de mots préalablement produits, après une interpolation avec le modèle *baseline* appris sur le corpus d'apprentissage ;
- en faisant varier le coefficient d'interpolation, on peut faire varier le taux d'erreur mots.

Avec cette méthode nous avons obtenu 4 décodages différents de notre corpus de test obtenus avec 4 valeurs différentes du coefficient d'interpolation (0,0 0,5 0,8 et 1.0). Ces décodages sont représentés par 4 séries de graphes de mots $G_{0,0}$, $G_{0,5}$, $G_{0,8}$ et $G_{1,0}$ dont les scores sont une combinaison des modèles acoustique et de langage. Les graphes $G_{0,0}$ correspondent au décodage *baseline* où aucune données de test n'est intégrée dans l'apprentissage. Les taux d'erreurs mots des meilleures hypothèses de ces graphes sont :

Graphes	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$
WER	32,2	27,2	24,1	18,5

Même si l'introduction de données de tests dans l'apprentissage génère forcément un biais, il est réduit du fait que cette introduction n'intervient que dans la deuxième passe de l'étape de reconnaissance : les erreurs et confusions acoustiques produites par le modèle *baseline* sont toujours présentes. Cependant ce sont bien évidemment les résultats obtenus sur les graphes $G_{0,0}$ qui sont les plus réalistes puisqu'ils sont produits sans introduction des données de tests. Les autres graphes ne servent qu'à observer la corrélation taux d'erreur mots et taux d'erreur concepts.

4. STRATÉGIE D'INTERPRÉTATION

Nous noterons C l'interprétation d'un message. C représente une séquence de concepts de base, tels que ceux définis dans le corpus MEDIA et présenté dans l'exemple de la table 1. Le décodage conceptuel consiste à chercher la chaîne de concepts $C = c_1, c_2, \ldots, c_k$ maximisant $P(C|A)$, A étant la séquence d'observations acoustiques. Trouver la meilleure séquence de concepts $\hat{C}$ exprimée par la séquence de mots W à partir de la séquence d'observations acoustiques A s'exprime par la formule suivante :

$$P(\hat{C}\hat{W}|A) \approx \max_{C,W} P(A|W)P(W,C) \qquad (1)$$

Les deux stratégies possibles pour obtenir $\hat{C}$ sont :

- chercher tout d'abord la meilleure chaîne de mots $\hat{W}$ étant donné A, puis chercher la meilleure séquence de concepts $\hat{C}$ sur la chaîne $\hat{W}$; nous appellerons cette approche l'approche *séquentielle* ;
- chercher conjointement la meilleure chaîne de mots et la meilleure séquence de concepts, tel que cela est exprimé dans l'équation 1 ; c'est l'approche *intégrée* proposée dans cette étude.

Cette recherche par l'approche intégrée de la meilleure interprétation $\hat{C}$ est faite dans un graphe de mots produit par le système de RAP pour chaque message traité. La première étape dans cette recherche consiste à transformer ce graphe de mots en un graphe de concepts. Le principe général de cette approche est décrit dans [4], son utilisation dans la campagne MEDIA est présentée dans [5], nous allons brièvement la résumer dans le paragraphe suivant.

Les constituants sémantiques sont appelés *tags conceptuels* et sont notés c. Ils correspondent aux 83 attributs présentés dans l'ontologie MEDIA (les informations sur les spécifieurs et les modes sont associés à un autre niveau d'interprétation dans notre système). À chaque tag c est associée la chaîne de mot W^c supportant le concept et à partir de laquelle sa valeur va être extraite, comme dans l'exemple de la table 1.

Dans le module de compréhension développé au LIA il existe un automate à états finis (ou Finite State Machine *FSM*) pour chacun de ces concepts. Ces automates sont

[1] téléchargeable à l'adresse :
http ://www.lia.univ-avignon.fr/chercheurs/bechet/

des transducteurs qui acceptent les séquences de mots W^c en entrée et qui produisent en sortie les concepts c correspondant. Ces transducteurs peuvent être créés manuellement pour les concepts indépendants du domaine (par exemple les dates ou les prix), ou induits par apprentissage pour les concepts propres au corpus MEDIA. L'ensemble de ces transducteurs est regroupé en un seul automate appelé *automate conceptuel*, auquel est ajouté un automate *filler* pour accepter tout ce qui ne fait pas partie d'un concept.

En effectuant une opération d'intersection entre le graphe de mots produit par le système de RAP et cet *automate conceptuel*, nous obtenons directement un transducteur où les chemins sur les symboles d'entrées sont des chaînes de mots et les chemins sur les symboles de sorties sont des chaînes de concepts. Afin d'évaluer les différentes analyses possibles en concepts d'une même chaîne de mots, un étiqueteur en concepts à base de HMM, lui aussi représenté sous la forme d'un automate, est composé avec le transducteur obtenu.

Le résultat du processus de décodage est une liste de n-meilleures interprétations appelée *N-Best Structurée*. Cette liste contient les meilleures interprétations du transducteur final structurées selon deux niveaux : le premier niveau correspond aux meilleures chaînes de concepts ; le deuxième niveau contient pour chaque séquence de concepts les meilleures valeurs trouvées dans le transducteur. La dernière étape du processus d'interprétation réside dans le module de décision, basé sur des classifieurs, choisissant une hypothèse dans cette liste de n-meilleure hypothèses. C'est à cette étape que le contexte du dialogue peut intervenir. Dans les expériences présentées dans le paragraphe 5, le module de décision est réduit au choix de l'hypothèse de probabilité maximale dans le transducteur de décodage. Cette stratégie d'interprétation est présentée à la figure 4.

Nous appellerons cette méthode l'approche *intégrée*, dans la mesure ou la recherche de la meilleure chaîne de mots et de la meilleure chaîne de concepts est simultanée. Pour comparer cette approche à l'approche séquentielle traditionnelle, nous avons également fait les mêmes expériences en réduisant le graphe de mots produit par SPEERAL à la chaîne de mots de probabilité maximale selon les modèles de RAP. La chaîne de traitement est par la suite identique.

5. EXPÉRIENCES

Les expériences présentées dans cette étude ont été menées sur le corpus MEDIA en considérant les 83 attributs présentés au paragraphe 2. Le mode et les 19 spécifieurs ne sont pas pris en compte ici, ils sont traités dans notre système par le module d'interprétation d'un énoncé en contexte, et ne relèvent donc pas du processus de décodage conceptuel présenté ici. Les performances sont mesurées par rapport au taux d'erreurs sur les paires attribut/valeur (appelé le *Concept Error Rate* ou *CER*, cette mesure est obtenue de manière similaire au WER en considérant les concepts à la place des mots). Un concept détecté est considéré comme correct uniquement si l'attribut du concept ainsi que sa valeur normalisée sont corrects d'après la référence.

Le tableau 2 présente les résultats des deux approches, *séquentielle* et *intégrée* sur plusieurs graphes de mots. Nous observons que dans tous les cas l'approche intégrée surpasse l'approche séquentielle, sauf bien sûr pour les transcriptions manuelles où le graphe de mots est réduit à une seule chaîne. Ainsi l'approche intégrée permet d'obtenir des performances similaires à ce que l'on obtiendrait en séquentiel avec un taux d'erreurs mots inférieur de 15% en relatif (cf écart entre $G_{0,0}$ et $G_{0,5}$).

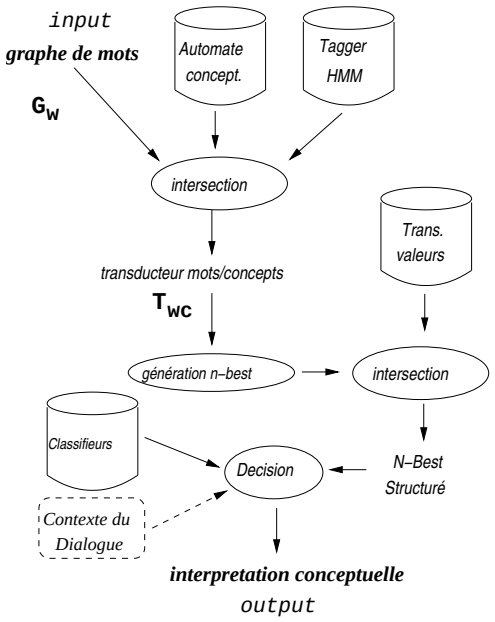


FIG. 1: Stratégie de compréhension du langage naturel oral du LIA

Il faut noter également que les performances obtenues sur les transcriptions manuelles sont comparables à celles des meilleurs systèmes ayant participés à l'évaluation MEDIA.

Un autre enseignement intéressant de ces expériences est la corrélation entre le taux d'erreur sur les mots (WER) et celui sur les concepts (CER). La figure 2 illustre cela en montrant une relation linéaire entre ces deux quantités.

Graphe	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$	Ref.
WER	32.2	27.2	24.1	18.5	0
CER (Seq.)	44.8	41.2	39.3	36.5	20.9
CER (Int.)	40.8	38.5	37.7	34.2	20.9

TAB. 2: WER et CER sur différents graphes avec l'approche séquentielle (Seq.) et l'approche intégrée (Int.). La colonne *Ref.* correspond au traitement de la transcription manuelle du corpus de test

La dernière expérience présentée dans cette étude concerne l'évaluation des listes de n-meilleures hypothèses produites par les différentes méthodes testées. Ces listes sont particulièrement intéressantes dans le cadre d'un dialogue car il est possible de fournir au gestionnaire de dialogue, non pas une hypothèse unique, mais plusieurs hypothèses que le contexte du dialogue peut aider à filtrer. Une mesure communément utilisée pour mesurer le potentiel d'un graphe ou d'une liste d'hypothèses est la mesure *Oracle*.

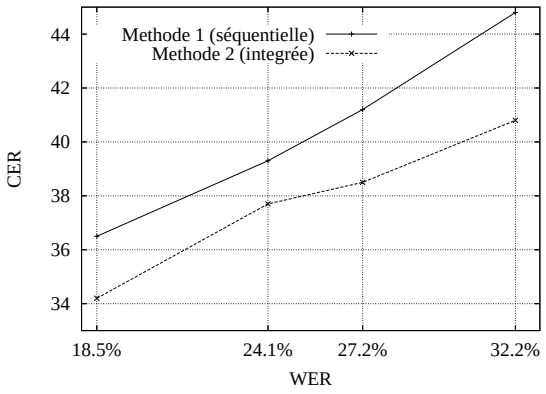


FIG. 2: Évolution du CER en fonction du WER

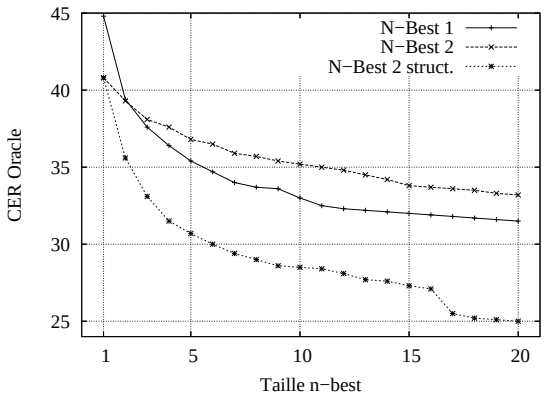


FIG. 3: Évolution du CER Oracle en fonction des tailles des listes de n-meilleures hypothèses pour deux méthodes : N-Best 1 = méthode séquentielle et N-Best 2 = méthode intégrée (avec et sans liste structurée)

Cette mesure consiste à sélectionner dans un ensemble d'hypothèses celle qui a le plus petit taux d'erreurs. Elle constitue ainsi le taux d'erreur minimal que ferait un système qui prendrait toujours la bonne décision sur le filtrage d'une liste d'hypothèses. Trois listes d'hypothèses ont été produites à partir des graphes $G_{0,0}$ (graphes *baseline* n'incluant pas de corpus de test dans l'apprentissage des modèles), leurs évaluations sont présentées sur la figure 3 :

- *N-Best 1* : cette liste est obtenue en énumérant la liste des n-meilleures hypothèses obtenues avec la méthode séquentielle ; la chaîne de mots étant fixe, chaque hypothèse diffère au niveau de la liste des concepts ;
- *N-Best 2* : cette fois les n-meilleures hypothèses sont les n-meilleures chemins dans le transducteur générées avec la méthode intégrée ; la chaîne de mot n'étant pas fixe, les n-meilleures chemins contiennent souvent la même suite de concepts et ne varie que par des choix de mots différents ;
- *N-Best 2 struct.* : correspond au N-Best Structuré décrit dans [4], et obtenu avec la méthode intégrée.

Comme le montre la figure 3, le N-Best Structuré permet d'éviter le principal inconvénient des listes de n-meilleures hypothèses produites à partir de graphes de mots : la génération d'hypothèses qui ne diffèrent que par des mots non signifiant du point de vue de l'interprétation du message.

En structurant cette liste par chaînes de concepts et valeurs, on obtient un résumé de toutes les interprétations possibles contenues dans le graphe en un nombre restreint d'hypothèses. Par exemple, en ne gardant que les 3 meilleures hypothèses du N-Best Structuré, on obtient le même taux Oracle qu'avec la liste complète des 20 meilleures hypothèses des autres méthodes.

6. CONCLUSION

Nous avons présenté dans cette étude un modèle de décodage conceptuel, basé sur une approche stochastique, intégré directement dans le processus de Reconnaissance Automatique de la Parole (RAP). L'un des principaux avantages de cette approche est de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Ainsi l'incertitude dans l'interprétation d'un message peut-elle être gardée plus longtemps pour être levée par des niveaux supérieurs d'interprétation intégrant le contexte du dialogue.

Les expériences menées sur le corpus MEDIA montrent que les performances du décodage conceptuel se dégradent linéairement en fonction du taux d'erreurs sur les mots. Nous avons cependant montré qu'une approche *intégrée* cherchant conjointement la meilleure séquence de mots et de concepts donnait de meilleurs résultats qu'une approche séquentielle.

Enfin la génération d'une liste de n-meilleures hypothèses structurées permet de réduire considérablement le nombre d'hypothèses susceptibles d'être envoyées au gestionnaire de dialogue, en gardant le même taux d'erreurs Oracle que la liste complète.

RÉFÉRENCES

[1] Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. Semantic annotation of the french media dialog corpus. In *Proceedings of Eurospeech*, Lisboa, Portugal, 2005.

[2] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1) :69–88, 2002.

[3] P. Nocera, G. Linares, and D. Massonie. Principes et performances du décodeur parole continue Speeral. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2002.

[4] Christian Raymond, Frédéric Béchet, Renato De Mori, and Géraldine Damnati. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48,3-4 :288–304, 2006.

[5] Christophe Servan and Frederic Bechet. Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. In *TALN*, Leuven, 2006.

Un modèle stochastique de compréhension de la parole à 2+1 niveaux

Hélène Bonneau-Maynard

LIMSI/CNRS
Groupe Traitement du Langage Parlé
helene.maynard@limsi.fr

Fabrice Lefèvre

LIA/Université d'Avignon
Equipe Dialogue Homme-Machine
fabrice.lefevre@univ-avignon.fr

ABSTRACT

In this paper an extension is presented for the 2-level stochastic speech understanding model, previously introduced in the context of the ARISE corpus [6]. In the new model, the additional stochastic level is in charge of the attribute value normalisation. Due to data sparseness, the full (3 level) model is not applicable straightforwardly and a variant is introduced where the conceptual decoding and value normalisation phases are decoupled.

The proposed approach is evaluated on the French MEDIA task (hotel booking and tourist information). This recent corpus has the advantage to be semantically annotated with conceptual segments, which allows for a direct training of the 2-level model. We also present some further model improvements such as the modality propagation or the 2-step hierarchical recomposition. On the whole, the various proposed techniques reduce the understanding error rate from 37.6% to 28.8% on the development set (24% relative improvement). This model has been engaged in the 2005 MEDIA evaluation campaign where it achieved the best results amoung the 5 participants with an error rate of 29%.

1. INTRODUCTION

Les approches stochastiques pour la compréhension de la parole offrent une alternative efficace aux approches par règles en réduisant le recours à l'expertise humaine et, ainsi, le coût global de développement du modèle [1, 2, 3, 4, 5]. Dans un précédent article [6], le développement d'un modèle stochastique a été présenté sur la tâche de dialogue ARISE (renseignements sur des horaires de trains et réservation de billets). Le présent article décrit les deux principales améliorations de notre modèle de compréhension par rapport au modèle de base : l'utilisation de corpus d'apprentissage segmenté sémantiquement (par rapport aux précédents travaux dans lesquels l'annotation sémantique portait sur des mots clés uniquement) et le recours aux modèles stochastique pour la normalisation des valeurs des attributs.

Avec une annotation sémantique par mots clés telle qu'elle était réalisée dans le corpus ARISE, la modélisation stochastique à 2 niveaux était fondée sur des segments sémantiques de taille fixe. Ces segments, centrés sur les mots clés précisés par l'annotation, étaient déterminés artificiellement a posteriori. Dans le nouveau schéma d'annotation, l'annotation sémantique est alignée sur des séquences de mots qui sont déterminés par les annotateurs humains. Les segments ont donc des tailles variables, ajustées en fonction des situations, et doivent permettre un meilleur apprentissage des modèles à 2 niveaux. Parallèlement, le modèle de compréhension est transformé en un modèle stochastique complet : la normalisation des valeurs est intégrée au processus stochastique, alors qu'elle était précédemment obtenue par le biais des règles semi-manuelles.

Notre participation au projet d'évaluation Technolangue EVALDA-MEDIA nous a permis de mettre au point et d'évaluer les améliorations proposées sur une nouvelle tâche. La tâche MEDIA concerne la réservation de chambres d'hôtel accompagnées de demande d'informations touristiques en France, les informations provenant d'une base de données disponible sur Internet.

L'organisation de l'article est la suivante : la prochaine section décrit la représentation sémantique. La modélisation stochastique intégrée est décrite dans la section 3. Finalement, après les descriptions du corpus MEDIA et des conditions expérimentales, la dernière section présente les résultats obtenus sur l'ensemble de développement et lors du test.

2. REPRÉSENTATION SÉMANTIQUE

La représentation sémantique du projet MÉDIA, décrite en détail dans [8], est fondée sur des structures d'attributs-valeurs dans lesquelles les relations hiérarchiques entre les concepts sont implicitement représentées par les noms et l'ordre des attributs. Chaque tour de parole est segmenté en un ou plusieurs segments sémantiques alignés sur les séquences de mots. Pour la compréhension littérale, un énoncé est représenté par une suite de segments sémantiques, chaque segment étant représenté par un triplet qui contient :
- le mode : affirmatif '+', négatif '-', interrogatif ' ?' ou optionnel '~' ;
- le nom de l'attribut représentant le sens de la séquence de mots ;
- la valeur de l'attribut.

L'ordre des triplets dans la représentation suit l'ordre des segments dans l'énoncé. Les valeurs des attributs sont des nombres, des noms propres ou des classes sémantiques qui regroupent des unités lexicales qui sont équivalentes pour la tâche. Des segments d'un même énoncé peuvent porter des modes différents.

La hiérarchie des attributs de base de la tâche est définie dans un dictionnaire sémantique. Différentes classes d'attributs y apparaissent. Certains attributs, dits **attributs BD**, (ex : `nom-hotel`) sont directement issus de la base de données liée à la tâche. Les attributs dits **mo-**

difieurs (e.g. `comparatif`), associés aux attributs BD, permettent d'en modifier le sens. Les attributs dits **généraux** correspondent aux commandes relatives à la tâche (`reservation`), ou au dialogue (`reponse`). Un des attributs généraux est utilisé pour représenter les références linguistiques (`lien-ref`). Le dictionnaire sémantique définit également pour chaque attribut l'ensemble des valeurs normalisées qui lui sont associées. Trois types de définitions de valeurs sont utilisées : par liste de valeurs (ex : `comparatif` avec les valeurs `inferieur`, `superieur`...), par expressions régulières (ex : dates), ou enfin sans restriction (ex : noms de clients).

Une représentation hiérarchique des connaissances, permettrait d'exprimer les relations complexes entre les constituants explicitement. Cependant, dans une approche orientée corpus, une représentation hiérarchique complique grandement l'annotation manuelle des données d'apprentissage. Pour tenir compte de cette difficulté, la représentation MEDIA, qui repose sur une représentation à un seul niveau, a été enrichie d'un ensemble de **spécifieurs** qui, combinés avec les noms des attributs BD et les modifieurs, permettent d'établir la relation au constituant principal. En reproduisant les relations représentées par les spécifieurs et en utilisant l'ordre des segments, il est alors possible de reconstruire une représentation arborescente à partir de la représentation à plat.

Dans nos travaux précédents, l'annotation sémantique était fondée sur des mots-clés : les attributs étaient associés uniquement aux mots qui déterminaient leur valeur. Dans le nouveau schéma d'annotation, la requête est découpée en segments sémantiques : les attributs sont maintenant associés à des *séquences de mots* - les segments - qui en désambiguïsent le mieux leur sens.

3. MODÉLISATION STOCHASTIQUE INTÉGRÉE

Le but de la compréhension stochastique est de déterminer la séquence de concepts $C = c_1 c_2 \dots c_N$ qui va représenter le sens de l'énoncé en posant l'hypothèse qu'il existe une correspondance séquentielle entre les concepts et les séquences de mots [1]. Soit $W = w_1 w_2 \dots w_N$ la séquence de mots de la phrase, le processus de compréhension recherche la séquence de concepts qui maximise la probabilité *a posteriori*, qui peut être écrite selon la formule de Bayes :

$$\hat{C} = \arg\max_C \Pr(C|W) = \arg\max_C \Pr(W|C)\Pr(C)$$

$\Pr(W|C)$ est estimé au moyen de probabilités n-grammes de mots connaissant le concept associé au mot i :

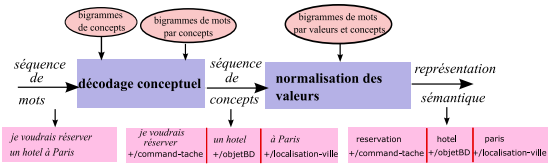
$$\Pr(W|C) \simeq \prod_{i=1}^{N} \Pr(w_i|w_{i-1}, \dots, w_{i-n}, c_i)$$

et $\Pr(C)$ est estimé par des probabilités m-grammes de concepts :

$$\Pr(C) \simeq \prod_{i=1}^{N} \Pr(c_i|c_{i-1}, \dots, c_{i-m})$$

A partir de cette formulation, plusieurs approches peuvent être considérées selon l'ordre des modèles utilisés pour produire l'estimation de $\Pr(W|C)$ et de $\Pr(C)$. Généralement, des bigrammes de concepts $\Pr(c_i|c_{i-1})$ ($m = 1$)

FIG. 1: Compréhension stochastique : modélisation à 2+1 niveaux.



sont suffisants pour modéliser les séquences de concepts. Lorsque l'on dispose d'une annotation sémantique segmentale pour le corpus d'apprentissage, on peut envisager d'utiliser des bigrammes de mots conditionnés au concept $\Pr(w_i|w_{i-1}, c_i)$ ($n = 1$). On parle alors d'une modélisation stochastique à 2 niveaux [6]. Afin d'améliorer la généralisation des modèles, il est possible d'utiliser un ensemble de classes lexicales.

La figure 1 représente les étapes du processus de compréhension. Dans notre modélisation, un concept est constitué de la combinaison du nom de l'attribut et de sa modalité. La première phase (*décodage conceptuel*) cherche à déterminer le concept le plus probable pour chaque sous-séquence de mots de l'énoncé. La seconde phase (*normalisation des valeurs*) consiste à déterminer, pour chaque attribut associé à chaque séquence de mots, la forme normalisée de la valeur attendue selon la représentation sémantique. Dans le schéma de la figure 1 par exemple, la séquence de mots *je voudrais réserver* associée à l'attribut `command-tache` lors de la phase de décodage conceptuel, doit être transformée en sa forme normalisée : `reservation`. Une même forme normalisée peut être produite par différentes séquences de mots : *je voudrais réserver, pour ma réservation....*

La normalisation est classiquement obtenue au moyen d'un ensemble de règles. Dans la modélisation décrite ici, nous proposons d'étendre la modélisation stochastique y compris à la phase de normalisation. Dans ce contexte, le modèle de compréhension peut être considéré comme un modèle intégrant 3 niveaux : mot, concept et valeur, comme indiqué dans les équations :

$$\begin{aligned} \hat{C}, \hat{V} &= \arg\max_{C,V} \Pr(C, V|W) \\ &= \arg\max_{C,V} \Pr(W|C, V)\Pr(C, V) \end{aligned}$$

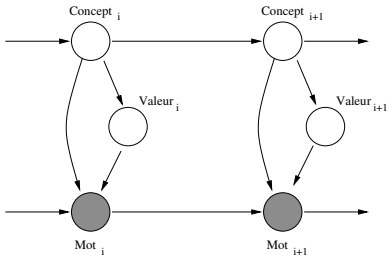
Cette modélisation conduit à des probabilités conditionnées par les valeurs normalisées et par conséquent à prendre en compte un nombre d'états considérablement augmenté. De plus, une telle solution est peu adaptée dans le cas où la liste des valeurs associées à un concept est ouverte (cas par exemple des nombres ou des noms de clients). Pour ces raisons, dans notre modèle, la normalisation n'a pas été totalement intégrée à la phase de décodage mais découplée de celui-ci pour être effectuée dans un second processus stochastique :

$$\hat{C} = \arg\max_C \sum_V \Pr(W|C, V)\Pr(C, V) \quad (1)$$

$$\begin{aligned} \hat{V} &= \arg\max_V \Pr(\hat{C}, V|W) \\ &= \arg\max_V \Pr(W|\hat{C}, V)\Pr(\hat{C}, V) \end{aligned} \quad (2)$$

L'équation 1 permet une meilleure généralisation du modèle conceptuel. Par ailleurs, l'hypothèse que la normali-

FIG. 2: Représentation du modèle stochastique de compréhension selon le formaliste des réseaux Bayesiens dynamiques.



sation des valeurs n'influence pas ou très peu la détermination des attributs paraît acceptable. Selon cette hypothèse, l'estimation des modèles stochastiques de normalisation peut être effectuée indépendemment de celle des modèles de concepts. Dans notre cas, un bigramme de mots est construit pour chaque couple (concept, valeur). Les indépendances conditionnelles sur les probabilités sont représentées dans le diagramme de la figure 2 dans le formalisme des réseaux Bayesiens dynamiques. Ainsi, le modèle envisagé à 3 niveaux est finalement assimilé à un modèle à 2+1 niveaux et le décodage depuis la séquence de mots jusqu'à la séquence de concepts associés à leurs valeurs normalisées s'effectue en deux temps, comme indiqué dans la figure 1. La forme finale obtenue est une suite de triplets [`mode`, `attribut`, `valeur`] comme attendu dans la représentation sémantique.

4. EXPÉRIENCES ET RÉSULTATS

Pour la compréhension littérale, le corpus MEDIA (tableau 1) consiste en une portion d'apprentissage de 10965 requêtes client, une portion de développement de 1009 requêtes, et un corpus de test de 3003 requêtes. Tous sont transcrits et annotés sémantiquement. Les 676 noms propres apparaissant dans le corpus correspondent essentiellement aux noms de villes (201) et d'hôtels (548) et sont très ambigus pour la tâche.

L'outil d'évaluation développé pour le projet MEDIA effectue un alignement entre deux représentations sémantiques afin de les comparer en terme de suppression, insertion et substitution. En mode *complet*, l'alignement est effectué sur tout le triplet [`mode`, `attribut`, `valeur`]. En mode *simplifié*, seuls deux modes (affirmatif et négatif) sont distingués (les modes ? et ~ étant projetés sur le mode +). Enfin, en mode *relâché*, les spécifieurs ne sont pas pris en compte dans l'identification des noms d'attributs.

4.1. Normalisation par règles (référence)

Les résultats décrits dans cette partie reposent sur la modélisation décrite dans [6]. Le modèle conceptuel à deux niveaux est appris sur les 11k énoncés du corpus d'apprentissage. Grâce à l'annotation segmentale, aucune transformation des annotations n'est nécessaire - comme l'utilisation de marqueurs de concepts - pour estimer les bigrammes de concepts. Un ensemble de classes lexicales est utilisé pour généraliser l'estimation des bigrammes. Les classes sont dérivées des attributs liés à la base de données et se limitent à des mots qui sont syntaxiquement et sémantiquement équivalents pour la tâche (par exemple `aéroport charles de gaulle`) peut apparaître sous différentes formes de surface (*Aéroport Charles de Gaulle*, *aéroport de Gaulle*...). Afin de résoudre les ambiguïtés liées au fait qu'une même forme de surface peut correspondre à différentes classes lexicales (par exemple un nom de ville correspond souvent aussi à un nom d'hôtel), les classes utilisées sont déterminées sélectivement selon le concept indiqué dans l'annotation.

TAB. 1: Caractéristiques principales des énoncés client des corpus d'apprentissage et de développement.

	appr.	dév.
nombre d'énoncés	10965	1009
nombre moyen de mots par énoncé	4.8	5.4
nombre de mots différents	2115	794
nombre d'attributs observés	29980	3125
nombre moyen d'attributs par énoncé	2.7	3.1
nombre d'attributs différents	144	106

L'annotation segmentale permet également de dériver du corpus d'adaptation un ensemble de règles de réécritures qui sont utilisées pour la phase de normalisation des valeurs. Une règle de réécriture est dérivée pour chaque observation d'un concept dans le corpus d'adaptation. Afin d'améliorer leur généralisation, les règles sont regroupées indépendemment du mode, ainsi que pour tous les attributs dont les noms ne différent que par les spécifieurs.

Le processus de compréhension est effectué après une transformation des énoncés dans laquelle les séquences de mots qui correspondent à des classes lexicales déterminées lors de l'apprentissage sont remplacées par le nom de la classe correspondante. Les mots vides de sens comme *euh, ah* sont également retirés de l'énoncé avant le décodage conceptuel. Les taux d'erreur de compréhension pour ce système de référence sont donnés dans la première ligne du tableau 2 : de 37,6% en mode complet à 23,0% en mode relâché et 20,8% pour les valeurs seules.

4.2. Normalisation stochastique des valeurs

La modélisation 2+1 avec une normalisation stochastique découplée, décrite dans la section 3, est ensuite substituée à la normalisation par règles du système de référence. La normalisation stochastique permet une amélioration relative des résultats de 6% par rapport à une normalisation par règles (deuxième ligne du tableau 2, `norm. stoch.`). Le gain relatif passe à 7.6% si on complète la normalisation stochastique avec un système de pénalités (troisième ligne du tableau 2, `norm. stoch.+`). Les pénalités sont appliquées en distinguant trois cas :

1. aucun mot de la séquence traitée n'a jamais été observé pour la valeur considérée ;
2. la forme normalisée apparaît telle quelle parmi les mots de la séquence traitée ;
3. tous les autres cas.

Les probabilités fournies par le module de normalisation stochastique reçoivent un malus dans le premier cas, un bonus dans le deuxième et ne sont pas modifiées dans le dernier.

Les résultats tendent à montrer que la normalisation stochastique représente une alternative efficace à la normalisation par règles. Elle permet une bonne généralisation

TAB. 2: Taux d'erreur de compréhension (%) sur le corpus de développement : complet, relâché (2 modes, sans spécifieurs) et valeurs uniquement. La colonne *#cpt* donne le nombre de concepts dans le modèle.

	#cpt	complet	relâché	valeurs
référence	390	37,6	23,0	20,8
norm. stoch.	390	36,9	21,8	19,6
norm. stoch.+	390	36,9	21,6	19,2
modes+	393	35,6	21,3	19,2
spécifieurs+	344	28,8	21,5	19,7
test (*officiel*)	344	29,0	21,6	19,7
test (*corrigé*)	344	27,7	20,4	18,5

en dépit du très petit nombre d'observations disponible pour chaque valeur de chaque concept. Afin d'augmenter la quantité de données par valeur, une normalisation partagée a été évaluée, comme dans le cas de la normalisation par règles. Si nos expériences ont montré que cette méthode permettait un petit gain avec un partage indépendant du spécifieur, ce gain disparaît après l'intégration des deux techniques présentées ci-après. Elle n'a donc pas été retenue dans le système complet.

4.3. Identification des modes

La différence importante entre les résultats en mode complet et en mode relâché indique que les confusions sur les modes sont la source de nombreuses erreurs. Une difficulté vient de ce qu'au cours de l'annotation manuelle du corpus, un mode positif a été affecté à tous les segments associés à l'attribut `null`, introduisant ainsi des discontinuités artificielles qui bloquent la propagation des modes lors de la phase de décodage conceptuel (limitée aux successions de 2 concepts par la modélisation en bigrammes). Une modification automatique du mode de tous les segments `null` situés entre deux segments de même mode non affirmatif a donc été réalisée. Comme le montre la troisième ligne du tableau 2 (`modes+`), cette simple propagation des modes a permis - en introduisant 3 concepts supplémentaires : `-/null`, `?/null` et `~/null` - d'améliorer les résultats d'un score relatif de 3,6% sur le taux de compréhension en mode *complet*.

4.4. Recomposition hiérarchique

Le modèle stochastique ne permet pas de gérer les dépendances hiérarchiques à long terme. Afin de prendre en compte les limites du modèle, l'identification des spécifieurs de concepts - qui portent les dépendances à long terme dans la représentation MEDIA - est transformée en une procédure en deux étapes.

La plupart des spécifieurs sont activés dans des contextes sémantiques particuliers portant les dépendances à long terme, mais peuvent être décrits en terme de présence de concepts de base dans l'énoncé. Par exemple, le spécifieur `reservation` apparaît presque exclusivement dans le cas où le concept `command-tache` avec la valeur `reservation` a été identifié pour un autre segment de l'énoncé. Ces contextes peuvent donc être retrouvés *après* la phase de décodage conceptuel. C'est ainsi que les modèles sont *dorénavant* appris avec des concepts sans de spécifieurs - réduisant ainsi le nombre de concepts du modèle de 390 à 344. Les spécifieurs sont déterminés dans un second temps par un ensemble de règles. Cette procédure en deux temps a permis une amélioration relative des résultats de 19% (ligne `specifieurs+` du tableau 2), avec une très faible détérioration dans la normalisation des valeurs (de 19,2 à 19,7%), qui peut s'expliquer par l'augmentation du nombre possible de valeurs par concept.

CONCLUSION

Dans cet article, nous avons proposé et évalué un modèle de compréhension de la parole stochastique à 2+1 niveaux. Par rapport au modèle de référence à 2 niveaux, une amélioration relative de 24% du taux d'erreur de compréhension a été obtenue avec un modèle comptant 344 concepts (incluant des constituants hiérarchiques). Une part importante de l'amélioration des performances provient d'une technique simple mais efficace de traiter la composition hiérarchique en 2 étapes. Toutefois, les erreurs sur les spécifieurs représentent toujours 25% du total des erreurs. Une amélioration de la méthode actuelle par l'introduction de classifieurs statistiques devrait permettre une meilleure couverture des contextes sémantiques pour une recomposition hiérarchique plus précise des concepts.

Le système développé a participé à l'évaluation MEDIA 2005. Les résultats sur l'ensemble de test (3003 énoncés utilisateurs) sont donnés dans la dernière partie du tableau 2. La première ligne correspond aux résultats officiels après adjudication, la seconde ligne correspond aux résultats obtenus après correction d'une erreur de manipulation lors de l'évaluation (échange de 2 fichiers). Avec un taux d'erreur de 29.0% en mode complet, le système se classe 1er parmi les 5 participants.

RÉFÉRENCES

[1] E. Levin and R. Pieraccini, "Concept-based Spontaneous Speech Understanding System," in ESCA Eurospeech, Madrid, 1995.

[2] R. Schwartz, S. Miller *et al.*, "Hidden Understanding Models for Statistical Sentence Understanding," in IEEE ICASSP, Munich, 1997.

[3] F. Pla , A. Molin *et al.*, "Language Understanding using Two-level Stochastic Models with POS and Semantic Units," LNCS series, vol. 2166, 2001.

[4] Y. He and S. Young, "Hidden Vector State Models for Hierarchical Semantic Parsing," in IEEE ICASSP, Hong Kong, 2003.

[5] C. Raymond, F. Bechet *et al.*, "On the use of finite state transducers for semantic interpretation," Speech Communication, vol. 48 :3-4, pp 288-304, 2006.

[6] F. Lefevre and H. Bonneau-Maynard, "Issues in the development of a stochastic speech understanding system," in ICSLP, Denver, 2002.

[7] L. Devillers, H. Maynard *et al.*, "The French MEDIA/EVALDA project : the evaluation of the understanding capability of Spoken Language Dialogue Systems," in LREC, Lisbon, 2004.

[8] H. Bonneau-Maynard, S. Rosset *et al.*, "Semantic annotation of the MEDIA corpus for spoken dialog," in ISCA Eurospeech, Lisbon, 2005.

Session XII

Poster

Mercredi 14 juin 2006 - 11h15 12h30

Evaluation de systèmes de génération de mouvements faciaux

O. Govokhina[(1,2)], G. Bailly[(2)], P. Bagshaw[(1)] , G. Breton[(1)] & R. Pastrana[(1)]

(1) France Telecom R&D, 4 rue du Clos Courtel, BP 59, F35512 Cesson-Sévigné Cedex

(2) Institut de la Communication Parlée, UMR CNRS 5009, INPG/U. Stendhal, 46, av. Félix Viallet - F38031 Grenoble

{ oxana.govokhina, paul.bagshaw, gaspard.breton, ricardo.pastrana}@francetelecom.com,
{oxana.govokhina,gerard.bailly}@icp.inpg.fr

ABSTRACT

This paper presents the implementation and evaluation of different movement generation techniques for speech-related facial movements. State-of-the-art systems are implemented. A novel system that combines HMM-driven pre-selection of diphones with a standard concatenation system is also implemented. The trajectory formation systems are parameterised using the same training material. The ground-truth data consists of facial motion and acoustic signals of one female speaker uttering 238 sentences. Both objective and subjective evaluation of the systems is reported. The objective evaluation observes the linear correlation coefficient between original and predicted movements. It is complemented by an audiovisual preference test where ground-truth and predicted movements drive a 3D virtual clone of the original speaker.

I. INTRODUCTION

La perception et la production de la parole sont bimodales. L'information complémentaire et redondante fournie par l'articulation acoustique et visuelle est utilisée efficacement par les interlocuteurs pour améliorer la détection de la parole [1] et son intelligibilité [2]. Les personnes sont très sensibles aux divergences audiovisuelles spatiales [3] et temporelles [4]. Les systèmes de synthèse de la parole qui peuvent aussi produire des signaux audiovisuels à partir des données phonologiques ou acoustiques doivent reproduire les co-variations observées dans la parole naturelle ainsi que contrôler la variabilité des gestes articulatoires.

La modélisation de la coarticulation est un problème difficile et non résolu [5]. La variabilité de l'articulation observée est largement planifiée [6] et exploitée par les interlocuteurs [7]. Depuis les premiers travaux d'Öhman sur la modélisation des mouvements linguaux [8], plusieurs modèles de coarticulation ont été proposés. Nous avons implémenté quelques modèles et nous les avons paramétrés et comparés avec des données de capture de mouvement.

Cet article est organisé comme suit: l'état de l'art est brièvement présenté dans la section 2, le corpus audiovisuel utilisé et la modélisation articulatoire sont décrits dans la section 3, les systèmes de synthèse visuelle implémentés sont dans la section 4, la méthodologie d'évaluation et les résultats sont présentés dans les sections 5, 6 et 7.

II. ETAT DE L'ART

Trois modules essentiels constituent un système d'animation faciale : contrôle, forme et apparence. Le module de contrôle calcule un ensemble de paramètres caractéristiques de la forme à partir d'une spécification de la chaîne phonétique à prononcer. Le module de forme se charge de calculer une géométrie à partir des paramètres caractéristiques calculés puis le modèle d'apparence se charge de texturer cette forme géométrique. Les modules de contrôle peuvent être divisés en deux catégories en fonction du type des données d'entrée [9]: ils sont calculés soit à partir d'un signal corrélé i.e. acoustique, soit depuis une spécification symbolique i.e. la chaîne phonétique. Les systèmes guidés par l'acoustique essaient de générer les mouvements faciaux qui auraient produit les sons correspondants. Un décodage phonémique intermédiaire n'est pas obligatoire [10, 11].

Les systèmes opérant à partir de la chaîne phonétique peuvent être grossièrement divisés en différentes catégories: systèmes basés visemes [12], systèmes basés coarticulation [13, 14], systèmes de modélisation de trajectoires [15, 16] et systèmes basés concaténation [17-19].

La comparaison de ces systèmes est problématique [9] car les modèles sont construits à partir des différents corpus et leurs méthodologies d'évaluation sont différentes. L'approche modulaire est rarement possible car les modèles de contrôle, de forme et d'apparence sont souvent interdépendants. La qualité du rendu influence aussi beaucoup la qualité perçue des modèles de contrôle [20]. Ici, l'objectif est d'évaluer les systèmes de synthèse visuelle existants et de proposer un nouveau modèle basé données qui profite des meilleures solutions existantes.

III. DONNEES AUDIOVISUELLES ET MODELISATION ARTICULATOIRE

Les systèmes de synthèse que nous allons évaluer sont construits à partir de données audiovisuelles. La base de données utilisée comprend 238 phrases du français prononcées par une locutrice. Les données acoustiques et vidéo sont capturées par un système Vicon© [21]. Le système capture, à 120 Hz, les positions 3D des 63 marqueurs infrarouges réfléchissants qui sont posés sur le visage de la lectrice (voir Figure 1). Le signal acoustique, à 11025 Hz, est segmenté semi-automatiquement en phonèmes.

Un modèle de forme est construit à partir des positions 3D des 63 points caractéristiques. La méthodologie du *clonage* développée à l'ICP [14, 22] consiste en une série de régression linéaire de composantes obtenues par Analyse en Composantes Principales (ACP) appliquée sur des sous-ensembles pertinents des points caractéristiques. D'abord, la contribution de la rotation de la mâchoire (*Jaw*1) est estimée et soustraite des données. Ensuite, Le mouvement d'arrondissement des lèvres (*Lips*1) est estimé à partir du résidu et soustrait des données. Les mouvements verticaux des lèvres du haut et du bas (*Lips*2 and *Lips*3), des coins des lèvres (*Lips*4), de l'avancée de la mâchoire (*Jaw*2) et de la gorge (*Lar*1) sont soustraits dans cet ordre des données résiduelles. Ainsi sept paramètres articulatoires sont obtenus. Leur contribution à l'explication la variance des mouvements est présentée dans le Tableau 1.

Un clone 3D est construit à partir d'enregistrements vidéo photogrammétriques de la même lectrice [23]. Le clone virtuel vidéo-réaliste (voir Figure 1) est contrôlé par les sept paramètres articulatoires déterminés ci-dessous.

Tableau 1. Contribution des paramètres articulatoires à la variance globale.

Paramètre	*Jaw*1	*Lips*1	*Lips*2	*Lips*3	*Lips*4	*Jaw*2	*Lar*1
Variance	18.84	15.93	15.10	14.28	14.07	11.91	9.88
Cumulée	18.84	34.77	49.87	64.15	78.22	90.13	100

Figure 1. Gauche: Disposition des points caractéristiques utilisés pendant la capture des mouvements. Droite: Le clone virtuel 3D de la lectrice.

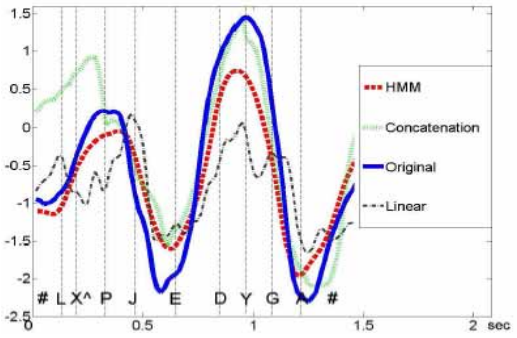


Figure 2. Trajectoires du paramètre Jaw1 de rotation de mâchoire générée par différents systèmes pour la phrase "Le pied du gars".

Les systèmes de génération des mouvements faciaux liés à la parole sont paramétrés par les données audiovisuelles de 228 phrases. Les dix phrases restantes sont utilisées dans le test et contiennent des diphones qui peuvent être retrouvés dans la base des données en au moins deux exemplaires.

IV. LES SYSTEMES DE SYNTHESE VISUELLE

Plusieurs modèles de synthèse représentant les approches décrites auparavant sont implémentés:

IV.1. Les modèles guidés par l'acoustique

Les paramètres articulatoires sont calculés directement à partir des paramètres acoustiques. Des signaux de 16.7 ms sont extraits à partir du signal d'origine en synchronie avec les données visuelles. Douze paramètres LSP (Line Spectrum Pair) et l'énergie sont calculés et lissés [24]. Enfin, un modèle de régression linéaire qui relie les paramètres acoustiques aux paramètres articulatoires est estimé. A la synthèse, les paramètres articulatoires sont générés à partir des paramètres acoustiques grâce au modèle obtenu. Notons que les systèmes de mise en correspondance non linéaires n'améliorent pas significativement les corrélations ni la qualité globale.

IV.2. Les modèles de synthèse basés HMM

Le principe de synthèse vocale par HMMs fut introduit par Donovan [25] et étendu à la synthèse audiovisuelle par le groupe HTS [16]. La technique de synthèse par HMMs comprend les étapes d'apprentissage et de synthèse. Son application à la synthèse visuelle est décrite ci-dessous.

Apprentissage. Un HMM et un modèle des durées d'états sont appris pour les paramètres articulatoires de chaque phonème (en contexte ou non) de la base d'apprentissage. Les vecteurs d'observation sont constitués des paramètres visuels statiques et dynamiques, c'est-à-dire, des valeurs des paramètres articulatoires et de leurs dérivées. L'estimation des paramètres des HMMs est basée sur le calcul de maximum de vraisemblance (*Maximum-Likelihood*) [26]. Cette estimation est effectuée par un algorithme spécifique de EM (*Expectation Maximisation*) connu comme l'algorithme récursif de Baum-Welch. Ainsi, un modèle gauche-droit à 3 états avec des distributions gaussiennes simples (covariance diagonale) est appris pour chaque phonème.

Synthèse. D'abord, une séquence de modèles HMM correspondants à la chaîne phonétique est construite. Les durées des états sont déterminées en répartissant la durée phonémique de manière adéquate [27]. Une fois la séquence d'états spécifiée, la trajectoire des paramètres articulatoires est estimée par résolution d'une équation linéaire [28]. Cet algorithme exploite la dépendance entre paramètres statiques et dynamiques. Ainsi, ce système est équipé théoriquement pour prendre en compte l'effet de coarticulation.

IV.3. Les modèle de synthèse par concaténation

La synthèse de la parole par concaténation consiste en la sélection et concaténation d'unités préenregistrées dans un dictionnaire. Tout d'abord des caractéristiques phonologiques sont utilisées pout sélectionner les unités candidates: la correspondance phonémique est exigée prioritairement mais des contraintes phonotactiques (contexte phonémique, position dans la syllabe ou dans le mot) et/ou phonologiques de plus haut niveau peuvent être ajoutées [29]. Ensuite, un algorithme de programmation dynamique trouve un chemin optimal à travers le treillis des candidats qui minimise un coût cumulé de sélection et de concaténation.

Le coût de sélection incorpore souvent des pénalités dues aux contraintes de sélection non respectées. Si un modèle prosodique est disponible, la déviation entre paramètres prosodiques calculés et sélectionnés est aussi souvent prise en compte [30]. Le coût de concaténation est calculé en fonction de la distance entre les unités adjacentes à la frontière. Les caractéristiques statiques et dynamiques sont souvent considérées.

Ici, les candidats sont les diphones. Aucun coût de sélection n'est considéré. Les coûts de concaténation sont égaux aux distances euclidiennes entre les paramètres articulatoires aux frontières des unités pondérées par la variance globale expliquée (voir Tableau 1).

Enfin, les trajectoires des unités sélectionnées sont élargies/compressées non linéairement pour correspondre aux durées des diphones puis un algorithme spécifique de lissage anticipatoire est appliqué [31].

IV.4. Le modèle de synthèse par concaténation basé HMM

Un nouveau modèle qui utilise la prédiction par HMMs pour présélectionner les candidats a été implémenté. Les diphones présélectionnés à la première étape du système de concaténation sont ensuite classés dans l'ordre décroissant du coefficient de corrélation entre les trajectoires des diphones

Mean correlation	Nat	Inv	Lin	HMM	Conc (N=3)
1	1,00	-1,00	0,17	**0,55**	0,50
2	1,00	-1,00	0,26	**0,63**	0,47
3	1,00	-1,00	0,26	**0,58**	0,30
4	1,00	-1,00	0,18	**0,70**	0,66
5	1,00	-1,00	0,41	0,56	**0,64**
6	1,00	-1,00	**0,62**	0,54	0,56
7	1,00	-1,00	0,12	**0,60**	0,41
8	1,00	-1,00	0,39	**0,55**	0,20
9	1,00	-1,00	0,33	0,49	**0,56**
10	1,00	-1,00	0,40	0,59	**0,67**
Global	1,00	-1,00	0,31	**0,58**	0,50

Tableau 2. Evaluation objective des modèles phrase par phrase.

Vote (%, Nb)	Nat	Inv	Lin	HMM	Conc (N=3)
1	14,30 (3)	4,80 (1)	0	14,30 (3)	**66,70 (14)**
2	33,30 (7)	0	0	28,60 (6)	**38,10 (8)**
3	19,00 (4)	0	0	**57,10 (12)**	23,80 (5)
4	28,60 (6)	0	0	**52,40 (11)**	19,00 (4)
5	**85,70 (18)**	0	0	9,50 (2)	4,80 (1)
6	0	0	0	38,10 (8)	**61,90 (13)**
7	**71,40 (15)**	0	0	28,60 (6)	0
8	**57,10 (12)**	0	0	38,10 (8)	4,80 (1)
9	**81,00 (17)**	0	0	14,30 (3)	4,80 (1)
10	38,10 (8)	0	0	**57,10 (12)**	4,80 (1)
Global	**42,9**	0,5	0	33,8	22,8

Tableau 3. Mean preference scores.

de la base des données et celles prédites par HMMs. Les N meilleurs candidats sont retenus dans le treillis pour la sélection finale du modèle de concaténation. Notons que N=∞ correspond au modèle de concaténation initial et qu'une méthode de sélection moins brutale aurait consisté à utiliser le coût de sélection pour pénaliser les segments les moins corrélés.

V. EVALUATION OBJECTIVE

Les modèles de synthèse visuelle proposés sont paramétrés à partir de la base d'apprentissage. Les dix phrases de test sont synthétisées. Le coefficient de corrélation linéaire (coefficient de Pearson) entre les trajectoires synthétiques et celles d'origine est utilisé pour l'évaluation objective.

Cette première évaluation est mise à profit pour paramétrer de manière optimale les systèmes.

Les corrélations moyennes dans le cas de la synthèse par HMMs augmentent si les paramètres dynamiques sont pris en compte pendant les phases d'apprentissage et de synthèse. La corrélation est significativement plus importante quand la dérivée première est utilisée. L'utilisation de la dérivée seconde n'augmente cette corrélation que de manière marginale.

La corrélation moyenne dans le cas de la synthèse par concaténation en fonction des différentes valeurs de N atteint une valeur optimale pour N=3.

VI. EVALUATION SUBJECTIVE

Le but du test subjectif utilisé est d'évaluer la préférence globale des modèles proposés par rapport aux mouvements faciaux d'origine. Il faut noter que cette référence – souvent absente dans l'ensemble des stimuli utilisés dans les tests publiés – est très importante [31, 32].

Les trajectoires articulatoires des dix phrases sont générées par trois modèles: (a) le système de synthèse basé HMM avec les vecteurs articulatoires comprenant la dérivée première (*HMM*); (b) le système de synthèse par concaténation avec la méthode de présélection proposée et N=3 (*Conc*); (c) le système de synthèse par modèle de régression linéaire (*Lin*). Cet ensemble est complété par les trajectoires originales (Nat) et leurs inverses (*Inv*, où les paramètres originaux sont multipliés par -1) de manière à fournir aux sujets une gamme assez large de qualité.

Un exemple de trajectoires générées est montré Figure 2. Les résultats de l'évaluation objective correspondant aux modèles retenus sont dans le Tableau 2. La corrélation moyenne est maximale pour la synthèse par HMMs. Dans le cas d'une phrase, la corrélation est plus importante pour le modèle acoustique linéaire que pour le modèle de concaténation.

Les paramètres articulatoires générés sont utilisés pour l'animation du clone virtuel de la lectrice (voir Figure 1). Le signal acoustique original est joué en synchronie avec les mouvements faciaux. Ici, le test de préférence moyenne (*Mean Preference Score: MPS*) est utilisé. Chaque participant doit alors choisir la séquence qu'il préfère parmi cinq pour chaque phrase. Les 21 sujets qui ont participé à l'expérience n'ont aucune pathologie audiovisuelle. Les sujets peuvent jouer les stimuli tant de fois qu'ils désirent et peuvent changer leurs choix. L'ordre initial des séquences pour chaque phrase est aléatoire. Le test est effectué dans un environnement de luminance contrôlé. Les conditions de la luminance de fond sont basées sur la ITU-R BT.500-9 (ITU-R, 1998).

VII. RESULTATS ET DISCUSSION

Les résultats du test subjectif sont présentés dans le Tableau 3. Le modèle le plus préféré est l'original (42.9%) suivi par le modèle HMM (33.8%) et le modèle de concaténation (22.9%). Les scores de préférence pour les modèles *Lin* et *Inv* sont très bas, 0% et 0.5% respectivement.

La méthode de synthèse par HMM est jugée comparable aux mouvements originaux; les phrases générées par HMM étant de plus toujours préférées par au moins deux personnes. La synthèse par concaténation guidée HMMs est moins performante mais les résultats dépendent des phrases. Il est intéressant de constater que les mouvements de synthèse (*HMM* ou *conc*) sont préférés aux originaux pour six des dix phrases. Cela peut provenir des imperfections des modèles de forme et d'apparence mais les mouvements générés par ces deux modèles de prédiction sont jugés globalement comme adéquats aux mouvements originaux.

Le modèle acoustique linéaire a le score le plus bas (voir aussi les résultats précédents obtenus par Gibert et al [31]) même si sa corrélation objective est parfois importante et même proche de celle obtenue par le modèle de concaténation pour certaines phrases.

VIII. CONCLUSIONS

Des différentes méthodes de synthèse visuelle sont évaluées objectivement et subjectivement. Une nouvelle méthode proposée concatène les segments articulatoires présélectionnés grâce à une méthode basée HMMs. L'utilisation de cette méthode augmente considérablement la corrélation entre les trajectoires synthétiques et originales. Ce gain ne permet pas cependant d'atteindre ceux de la synthèse

purement HMM. Dans l'ensemble, les résultats de l'évaluation objective sont confirmés par l'évaluation subjective. Le système HMM semble être le plus efficace et le mieux accepté.

L'étude des résultats montre cependant que les résultats des évaluations dépendent du contenu phonétique des phrases. Le modèle HMM, s'il est meilleur en moyenne partout, génère des trajectoires moins coarticulées que celles produites par le système par concaténation. C'est dans cet esprit que nous avions décidé de coupler la solide charpente construite par HMM avec la richesse des détails phonétiques capturés par la synthèse par concaténation. Nous allons continuer à suivre cette idée qui devrait à terme produire un système à la fois robuste et fin.

BIBLIOGRAPHIE

[1] K. W. Grant and P. F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America*, vol. 108, pp. 1197-1208, 2000.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[4] N. F. Dixon and L. Spitz, "The detection of audiovisual desynchrony," *Perception*, vol. 9, pp. 719-721, 1980.

[5] W. J. Hardcastle and N. Hewlett, *Coarticulation: Theory, Data, and Techniques*. Cambridge, UK: Press Syndicate of the University of Cambridge, 1999.

[6] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, pp. 3-35, 1990.

[7] K. G. Munhall and Y. Tohkura, "Audiovisual gating and the time course of speech perception," *Journal of the Acoustical Society of America*, vol. 104, pp. 530-539, 1998.

[8] S. E. G. Öhman, "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, vol. 41, pp. 310-320, 1967.

[9] J. Beskow, "Talking heads. Models and applications for multimodal speech synthesis." Stockholm: KTH, 2003, pp. 63.

[10] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," presented at 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany, 2000.

[11] S. Curinga, F. Lavagetto, and F. Vignoli, "Lips movements synthesis using time-delay neural networks," presented at EUSIPCO, Trieste - Italy, 1996.

[12] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, pp. 45-57, 2000.

[13] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. Tokyo: Springer-Verlag, 1993, pp. 141-155.

[14] L. Revéret, G. Bailly, and P. Badin, "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," presented at International Conference on Speech and Language Processing, Beijing - China, 2000.

[15] T. Okadome, T. Kaburagi, and M. Honda, "Articulatory movement formation by kinematic triphone model," presented at IEEE International Conference on Systems Man and Cybernetics, Tokyo, Japan, 1999.

[16] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," presented at EUROSPEECH, Budapest, Hungary, 1999.

[17] S. Minnis and A. P. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," presented at International Conference on Speech and Language Processing, Beijing, China, 1998.

[18] O. Engwall, "Evaluation of a system for concatenative articulatory visual speech synthesis," presented at International Conference on Speech and Language Processing, Boulder - Colorado, 2002.

[19] F. J. Huang, H. P. Graf, and E. Cosatto, "Triphone-based unit selection for concatenative visual speech synthesis," presented at International Conference on Acoustics, Speech and Signal Processing, Orlando, FL, 2002.

[20] I. Pandzic, J. Ostermann, and D. Millen, "Users evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330-340, 1999.

[21] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech," *Journal of Acoustical Society of America*, vol. 118, pp. 1144-1153, 2005.

[22] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533-553, 2002.

[23] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, pp. 331-346, 2003.

[24] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23-43, 1998.

[25] R. Donovan, "Trainable speech synthesis," in *Univ. Eng. Dept.* Cambridge, UK: University of Cambridge, 1996, pp. 164.

[26] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[27] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," presented at International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[28] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into hmm-based speech synthesis," presented at ISCA Speech Synthesis Workshop, Pittsburgh, PE, 2004.

[29] P. Taylor and A. W. Black, "Speech synthesis by phonological structure matching," presented at EuroSpeech, Budapest, Hungary, 1999.

[30] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," presented at International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, 1996.

[31] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," presented at IEEE Workshop on Speech Synthesis, Santa Monica, CA, 2002.

[32] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," Massachusetts Institute of Technology, Cambridge, MA, CBCL Paper #224/AI Memo #2003-003 February 2003.

Contraintes globales pour la sélection des unités en synthèse vocale

Adrian Popescu[1], Cédric Boidin[2], Didier Cadic[2]

[1] Département LUSSI, ENST Bretagne, Brest, France
[2] France Télécom, Division R&D, TECH/SSTP/VMI, Lannion, France
adrian.popescu@enst-bretagne.fr, {cedric.boidin, didier.cadic}@francetelecom.com

ABSTRACT

This work proposes an alternative unit selection method for corpus-based voice synthesis. It introduces the need of long term constraints in the cost function, which cannot be handled by the traditional Viterbi algorithm. Therefore another optimization algorithm, the simulated annealing, has been chosen for our experiments. It has been evaluated on a cost function encouraging long term F0 continuity. Although the results of our experiments do not show real improvement of the overall quality, they are a starting point for further research on this relevant issue.

1. INTRODUCTION

Les systèmes de synthèse par corpus ont amélioré significativement la qualité de la synthèse vocale. Leur succès est basé sur l'utilisation de grandes bases de données, associée à des algorithmes efficaces de sélection des unités. L'étape de sélection consiste à choisir la meilleure suite d'unités parmi toutes celles présentes dans le corpus.

Pour cela, on minimise une fonction de coût mesurant la fluidité du signal de parole synthétisé ainsi que son adéquation aux cibles issues des traitements linguistiques. Elle est généralement définie comme une somme pondérée de coûts-cibles et de coûts de concaténation [1], puis minimisée de manière optimale grâce à un algorithme de programmation dynamique. Ces choix ont prouvé leur efficacité et donnent de bons résultats pour la synthèse de phrases neutres. D'autres algorithmes d'optimisation ont été expérimentés, comme par exemple les algorithmes génétiques [2].

Cependant la forme de la fonction de coût limite le type de contraintes qu'il est possible de prendre en considération : elles ne peuvent porter que sur des unités prises isolément (grâce au coût-cible), ou bien sur des couples d'unités consécutives (coût de concaténation).

Cet article présente plus en détail les algorithmes de sélection actuels et leurs limites, propose alors une nouvelle fonction de coût qui inclut des contraintes à plus long terme, ainsi qu'un algorithme permettant d'intégrer cette fonction de coût : le recuit simulé. Il décrit ensuite les tests effectués et leurs résultats, engage une discussion sur ces résultats puis conclut.

2. SÉLECTION DES UNITÉS

La fonction de coût est généralement définie comme la somme pondérée de coûts-cibles et de coûts de concaténation [1], comme dans l'équation (1).

$$C_L(s) = w_C \sum_{k=2}^{n} C_C(u_{k-1}, u_k) + w_T \sum_{k=1}^{n} C_T(u_k) \qquad (1)$$

où s désigne la séquence des unités $(u_1, u_2, \ldots, u_n)$, $C_L(s)$ le coût total associé à cette séquence, $C_C(u_{k-1}, u_k)$ le coût de concaténation entre les unités u_{k-1} et u_k, $C_T(u_k)$ le coût-cible associé à l'unité u_k et enfin w_C et w_T les pondérations associées.

L'utilisation d'une telle fonction de coût est motivée par les contraintes suivantes : les unités doivent être choisies dans un contexte prosodique et linguistique adéquat (coût-cible) et les transitions entre unités consécutives doivent être fluides (coût de concaténation). Une telle fonction de coût permet également une réduction drastique de la complexité algorithmique.

En effet, pour une séquence de N unités, chacune représentée par M occurrences, le nombre total de combinaisons est M^N ; mais la minimisation d'une fonction de cette forme peut être effectuée avec une complexité réduite à $N \times M^2$ par un algorithme de programmation dynamique, généralement de type Viterbi [3]. Le gain de complexité est important mais, en contrepartie, les contraintes ne peuvent porter que sur une ou deux unités consécutives. Ceci semble cependant suffisant pour synthétiser des phrases neutres de bonne qualité.

3. CONTRAINTES GLOBALES

L'introduction de contraintes globales dans les algorithmes de sélection des unités semble nécessaire pour augmenter le contrôle de la sélection.

Par exemple, comme l'explique Hirai [4], la fonction de coût standard ne permet pas de suivre au mieux un contour de F0 (fréquence fondamentale) dont tous les paramètres ne seraient pas figés, comme par exemple sa composante continue (ou baseline) que l'on peut choisir de ne pas imposer. Il a proposé de résoudre le problème en effectuant plusieurs sélections pour

différentes valeurs des paramètres, puis en choisissant la séquence d'unités au meilleur coût global.

Nous proposons d'aborder différemment ce problème, en introduisant une nouvelle fonction de coût C_N :

$$C_N(s) = w_L C_L(s) + w_G C_G(s) \quad (2)$$

Le terme $C_L(s)$ correspond aux contraintes locales définies dans l'équation (1). Le terme $C_G(s)$ représente la nouvelle contrainte globale, intégrant des relations entre unités non adjacentes. w_L et w_G sont des pondérations.

Avec l'introduction du terme global, il est par exemple possible de suivre au mieux un contour de F0 indépendamment de sa composante continue : le terme C_G est alors égal à l'écart quadratique moyen entre le contour réel de la séquence courante et le contour-cible, ces deux contours étant auparavant centrés sur leurs moyennes.

Un autre exemple de contrainte globale est la volonté d'assurer une certaine continuité de F0 entre unités non adjacentes, plus particulièrement autour des régions non-voisées. Cette volonté ne peut pas être prise en compte par le coût de concaténation habituel d'ordre 1.

4. LE RECUIT SIMULÉ

L'introduction de contraintes globales augmente la complexité de la sélection, et nous ramène à un problème de dimension M^N.

Plusieurs algorithmes peuvent être utilisés pour résoudre ce problème, mais la solution optimale est, dans le cas général, hors de portée.

Dans cette étude nous utilisons l'algorithme du recuit simulé [5, 6] décrit en figure 1.

```
Seq_Courante      = Seq_Initiale;
Temp_Courante  = Temp_Initiale;
DO
      Seq_Perturbée     = PERTURB (Seq_Courante);
      Seq_Courante      = ACCEPT (Seq_Perturbée,
                  Seq_Courante, Temp_Courante);
      Temp_Courante  = UPDATE (Temp_Courante);
WHILE NOT (Critère_Arrêt);
RETOUR (Seq_Courante);
```

Figure 1 : Pseudo-code du recuit simulé.

Les variables utilisées dans la figure 1 sont les séquences d'unités et la température. La séquence initiale est choisie aléatoirement. Les trois fonctions utilisées dans l'algorithme sont PERTURB, ACCEPT, UPDATE :

PERTURB : cette fonction effectue une modification de la séquence d'unités courante pour obtenir une séquence perturbée. Dans notre cas, elle consiste à remplacer un certain nombre d'unités adjacentes par d'autres unités candidates aléatoirement choisies. Cette fonction de perturbation est très simple ; elle pourra par la suite être remplacée par des heuristiques plus complexes de choix des nouvelles unités guidant efficacement l'exploration.

ACCEPT : cette fonction accepte ou non la perturbation. La perturbation est acceptée avec une probabilité *P* donnée par l'équation (3).

$$P = \begin{cases} \exp\left(-\dfrac{C(s_p) - C(s_c)}{T}\right) & if \quad C(s_p) \geq C(s_c) \\ 1 & if \quad C(s_p) < C(s_c) \end{cases} \quad (3)$$

T est la température courante, $C(s_c)$ est le coût associé à la séquence courante, $C(s_p)$ celui associé à la séquence perturbée.

Ainsi, toutes les perturbations associées à une baisse de coût sont acceptées et celles associées à une augmentation du coût sont acceptées avec une probabilité dépendant de la hausse du coût et de la température.

UPDATE : cette fonction met à jour la température. Deux lois sont généralement proposées pour la descente de température : la loi logarithmique et la loi géométrique. La première assure la convergence vers l'optimum global mais, pour des raisons de temps de calcul, nous avons choisi la loi géométrique, plus rapide.

Une fois les trois fonctions PERTURB, ACCEPT et UPDATE définies, les paramètres suivants doivent être fixés : température initiale, température finale, ainsi que la raison de la loi géométrique.

Le choix des paramètres est important pour assurer un bon comportement de l'algorithme. A titre d'exemple, si la température reste trop élevée, les perturbations sont trop souvent acceptées et aucune convergence intéressante ne peut être constatée ; si la température reste trop faible, les augmentations de coût sont rarement acceptées et on s'enferme rapidement dans un minimum local. Il faut donc diminuer la température lentement et dans un intervalle approprié afin d'explorer suffisamment de possibilités tout en assurant une convergence convenable. Dans notre cas ces paramètres sont fixés empiriquement.

5. COMPARAISON OBJECTIVE

5.1. Conditions de test

Cette partie vise à comparer les méthodes de sélection de façon objective. Le corpus de référence (7 heures de phrases du "Monde" enregistrées par une voix d'homme à 16 kHz), les fonctions de coût-cible et de coût de concaténation sont celles mises en œuvre dans

le système de synthèse de parole de France Télécom. Ces coûts sont basés sur des paramètres acoustiques ainsi que sur des étiquettes linguistiques et prosodiques issues des niveaux linguistiques du moteur de synthèse. Les unités de base sont les diphones. Aucun algorithme de traitement du signal n'est appliqué sur les unités, excepté un overlap-add lors de leur concaténation. Pour chaque diphone-cible, les 100 meilleures unités candidates sont présélectionnées en fonction de leur coût-cible, quelle que soit la méthode de sélection utilisée par la suite.

Un premier ensemble de 75 phrases journalistiques aléatoirement choisies dans "le Monde" représente le corpus de test A.

5.2. Comparaison algorithme de Viterbi - Recuit simulé

Une première expérience vise à mesurer les performances du recuit simulé pour la minimisation de la fonction de coût C_L.

Plusieurs variantes sont implémentées pour la fonction PERTURB, qui consiste à modifier un sous-ensemble d'unités adjacentes de la séquence courante. Dans notre cas le sous-ensemble et les unités de remplacement sont choisis aléatoirement. Plusieurs tailles de sous-ensemble sont testées : 1 unité, 2 unités adjacentes, 3 unités adjacentes, ou encore une taille aléatoire comprise entre 1 et 5. Il est à noter que les unités remplacées sont adjacentes dans la séquence courante, mais pas nécessairement dans le corpus de référence.

La table 1 montre le surcoût moyen de la solution fournie par le recuit simulé par rapport à la solution optimale (*i.e.* celle du Viterbi), calculé sur le corpus A.

Table 1 : Surcoût moyen de la solution du recuit simulé par rapport à la solution optimale (Viterbi), pour plusieurs tailles de fenêtres remplacées.

Nombre d'unités remplacées	Hausse de coût
1	15%
2	10%
3	14%
Aléatoire entre 1et 5	13%

On choisit donc pour la suite des expériences une fenêtre de modification de taille 2 unités, induisant un surcoût moyen de 10% par rapport au coût minimal. A titre de comparaison, la distribution des coûts de toutes les séquences possibles a une moyenne et un écart-type respectivement égaux à 270% et 35% fois le coût optimal. Le test d'écoute présenté au paragraphe 6.2 mesure l'impact perceptif de ce surcoût.

L'utilisation du recuit simulé augmente également le temps de calcul. Typiquement, la sélection des unités pour une phrase prend 10 secondes avec le recuit simulé contre 1 seconde avec l'algorithme de Viterbi.

6. EXPÉRIENCES ET RÉSULTATS

6.1. Exemple de contrainte globale

Nous avons ensuite appliqué l'algorithme du recuit simulé à la fonction de coût C_N de l'équation (2), en y intégrant la contrainte globale suivante :

$$C_G(s) = \sum_{i=2}^{Nv} B\left(\left| f_0^i - f_0^{i-1} \right|\right) \quad (4)$$

avec :

$$\mathrm{B}(x) = \begin{cases} x & if \quad x \geq Seuil \\ 0 & if \quad x < Seuil \end{cases} \quad (5)$$

f_0^i désigne la fréquence fondamentale du *i*-ième noyau vocalique et *Nv* le nombre de noyaux vocaliques de la phrase.

La contrainte globale décrite ci-dessus pénalise les écarts de F0 entre noyaux vocaliques adjacents, en particulier autour des régions non-voisées, ce qui ne peut être intégré dans le coût de concaténation.

Le *Seuil* est fixé à 5 demi-tons. Ce seuil est égal à celui employé par Mertens [7] pour caractériser l'intervalle primaire. Cet intervalle a été défini en lien avec le seuil du glissando caractérisant la variation minimale de fréquence perceptible.

6.2. Test d'écoute

Un deuxième ensemble de 30 phrases journalistiques aléatoirement choisies dans "le Monde" a été construit pour le test d'écoute ; c'est le corpus de test B. 15 auditeurs naïfs ont participé au test.

Chaque auditeur est invité à noter son impression globale pour chacune des 30 phrases sur une échelle à 5 niveaux, suivant la recommandation P.800 de l'UIT, où 1 correspond à "mauvais" et 5 à "excellent".

Les résultats du test sont présentés dans la table 2.

Table 2 : Résultats du test d'écoute sur les phrases du corpus B.

Corpus B	MOS
Algorithme de Viterbi, C_L	3.7
Recuit simulé, C_L	3.46
Recuit simulé, C_N	3.35

L'algorithme de Viterbi donne de meilleurs résultats MOS que le recuit simulé pour la fonction de coût C_L traditionnelle, ce qui indique que le surcoût de 10% est audible.

La dernière ligne de la table donne les résultats MOS pour la nouvelle fonction de coût C_N intégrant la

contrainte globale. Les résultats sont un peu inférieurs à ceux obtenus pour la fonction de coût standard C_L, toujours avec le recuit simulé, ce qui semble discréditer la contrainte globale choisie pour cette expérience.

7. DISCUSSION

La présente étude aborde les limites de la synthèse par corpus et propose une méthode de sélection des unités alternative qui utilise une fonction de coût différente et un algorithme d'optimisation associé. Les contraintes globales utilisées et l'algorithme d'optimisation sont cependant loin d'être au point.

En effet, la contrainte globale choisie ici est trop restrictive, décourageant toutes les discontinuités de F0 supérieures à 5 demi-tons. Les discontinuités fortes de F0 ont une signification linguistique en parole spontanée. Il faudrait donc, sur la base d'informations linguistiques, encourager certaines de ces discontinuités et en décourager d'autres.

L'utilisation du recuit simulé pour résoudre ce problème complexe est pratique puisque flexible et relativement indépendante des contraintes globales que nous souhaitons imposer. On pourrait cependant choisir d'autres algorithmes, ou définir une fonction de perturbation plus élaborée, qui substituerait intelligemment les unités de manière à faire converger l'algorithme plus rapidement.

8. CONCLUSIONS

Cette étude propose une méthode alternative de sélection des unités. Nous insistons sur la nécessité d'introduire de nouvelles contraintes dans la fonction de coût, à portée plus large que celles actuellement traitées par les algorithmes de sélection. La forme et la portée de ces contraintes imposent le choix d'un nouvel algorithme d'optimisation. Pour nos expériences, nous avons adopté le recuit simulé, qui donne des séquences d'unités proches en qualité de celles du Viterbi, bien que sous-optimales. Nous avons évalué ce nouvel algorithme de sélection sur une contrainte globale encourageant la continuité de F0 à long terme. Les résultats montrent que cette méthode détériore légèrement la qualité moyenne des phrases synthétisées, la contrainte choisie restreignant indistinctement toutes les variations de pitch, naturelles ou non. Les efforts doivent être poursuivis pour mettre au point un catalogue de contraintes globales plus efficaces ainsi que des heuristiques capables de guider intelligemment les perturbations en fonction des contraintes imposées et ainsi accélérer la convergence.

BIBLIOGRAPHIE

[1] A.W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis", *Proceedings of Eurospeech*, Rhodes, Greece, September 1995.

[2] R. Kumar, "A Genetic Algorithm for Unit Selection based Speech Synthesis", *Proceedings of ICSLP*, 2004.

[3] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. Inf. Th.*, vol. IT13, pp. 260-269, 1967.

[4] T. Hirai, S. Tenpaku, and K. Sikano, "Speech Unit Selection based on Target Values driver By Speech Data In Concatenative Speech Synthesis", *IEEE Workshop on Speech synthesis*, 2002.

[5] P.Y. Le Meur, "Synthèse de la parole par unités de taille variable", PhD. Thesis, 1996.

[6] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing", *Science*, Number 4598, 13 May 1983.

[7] P. Mertens, "Automatic recognition of Intonation in French and Dutch", *Proceedings of Eurospeech*, Vol. 1, pp. 46-50, 1989.

Une synthèse vocale destinée aux déficients visuels

Hélène Collavizza, Jean-Paul Stromboni

Laboratoire I3S, Ecole Polytechnique Universitaire de Nice Sophia Antipolis
helen@polytech.unice.fr , strombon@polytech.unice.fr

ABSTRACT

This paper presents an experiment in developping and testing a text to speech system which is needed for improving some applications dedicated to visually impaired users. When carrying out such applications, it appears that an interface able to speak, i.e. read a text, is mandatory. In order to allow an easy release, this text to speech system should be portable, licence free, using freewares and free solutions.

To fulfill all these needs, a solution was chosen and developped. On the one hand, several experiments were conducted with different such applications. On the other hand, the obtained text to speech system was made available on the Web for evaluation purpose. These two information sources allow to gather a set of advices and knowledge on such tools, and make possible and easier to develop next versions.

1. INTRODUCTION : NOTRE BESOIN

Depuis plusieurs années, l'Ecole Polytechnique Universitaire de Nice Sophia Antipolis organise des journées intitulées Déficients Visuels et Nouvelles Technologies (DeViNT). L'objectif est de mettre en présence des associations, des chercheurs, des entreprises, des organismes de formation spécialisés, et des déficients visuels (voir http://www.polytech.unice.fr/devint).

Cette manifestation est l'occasion de développer des applications logicielles dédiées aux déficients visuels à la fois pour les aider à utiliser l'informatique et ses outils, et aussi pour sensibiliser non seulement les élèves mais aussi les enseignants à cette problématique.

Les applications sont diverses, en particulier des jeux, et souvent dédiés aux enfants, d'autant plus que la collaboration avec des établissements spécialisés locaux, tel l'institut d'éducation sensorielle Clément Ader et l'école du Château à Nice est de plus en plus importante. Cette collaboration intervient pour la conception, le développement, le test et la diffusion des résultats.

De cette collaboration comme de nos premières expériences, il résulte que la fonction de synthèse vocale (ou TTS pour Text To Speech) est essentielle dans ce contexte. L'utilisation de la voix est l'un des moyens de compenser la cécité lorsqu'il faut lire. Par exemple, les jeux de mémoire peuvent être adaptés en demandant de reconnaître des cris d'animaux plutôt que des images. Dans ce cas, il suffit d'enregistrer des sons. Cependant, d'autres applications, par exemple pour l'apprentissage de la lecture, requièrent un TTS pour résoudre le problème d'accessibilité, puisque le texte à lire à voix haute n'est pas prédéfini. La synthèse vocale recherchée doit satisfaire certaines contraintes :

- la langue parlée doit être le français
- la synthèse doit être facilement utilisable sur un maximum d'applications
- les applications doivent être portables, en particulier sur les plateformes Windows qui constituent l'environnement de travail le plus utilisé par les déficients visuels
- on doit pouvoir les diffuser gratuitement.

Les logiciels les plus utilisés par les aveugles sont les logiciels JAWS et Supernova. Ces logiciels intègrent une synthèse vocale qui lit à voix haute le texte de la fenêtre active sur l'environnement de travail. De nombreuses autres synthèses vocales existent, commerciales comme Acapela, Speechissimo... ou libres comme le projet Oralux sous linux http://oralux.org. Or, ces synthèses vocales correspondent mal à l'une ou l'autre des contraintes ci-dessus. D'où la nécessité de construire l'outil adapté pour l'inclure sans obligations ni pénalités dans les interfaces des applications de la journée DeViNT.

Dans les pages suivantes, (1) nous énumérons les choix qui ont dû être faits, (2) nous décrivons la solution obtenue, et (3) nous présentons quelques uns des retours d'expérience des utilisateurs de la synthèse vocale.

2. NOS CHOIX DE DÉVELOPPEMENT

Pour construire un TTS adapté à nos besoins, plusieurs choix doivent être faits, tels le langage de développement, en l'occurence java, pour être utilisé aisément par nos étudiants de première année et pour faciliter la mise en œuvre des interfaces graphiques. En outre, il faut choisir un moteur TTS capable de lire un texte à voix haute. Ici le moteur MBROLA [1] de l'Université de Mons présente les avantages suivants :

- MBROLA est distribué gratuitement sous licence (utilisation non militaire ni commerciale)
- il peut être piloté en java, sous Windows ou Linux
- il fournit un ensemble de voix françaises
- le projet MBROLA coordonne les expériences et les initiatives d'une communauté scientifique autour de la synthèse vocale, comme le projet EULER [2] ou FipsVox [3], et de l'utilisation du moteur MBROLA.

Nous avons donc choisi d'apporter une contribution à ce projet en développant un TTS français en java basé sur MBROLA, ce qui n'existait pas alors.

2.1 La chaîne de synthèse MBROLA

Pour créer une synthèse vocale en utilisant MBROLA, il suffit en fait d'écrire un module de transcription de texte en phonèmes. En effet, le processus de synthèse vocale est scindé en deux blocs, nommés « Natural Language Processing » et « Digital Signal Processing » [4]. Le premier bloc est chargé du traitement d'un texte qu'il faut traduire en une séquence de phonèmes. Le deuxième bloc est l'outil MBROLA lui-même qui prend les phonèmes et d'autres paramètres comme la voix à utiliser, pour lire le texte à voix haute et l'enregistrer dans un fichier audio (au format wave). Un constat important est que MBROLA rassemble et résout les problèmes de Traitement du Signal. Le travail restant, la construction du bloc « Natural Language Processing », est plus motivant pour nos étudiants en sciences informatiques.

2.2 Le bloc « Natural Language Processing»

Le bloc « Natural Language Processing » prend un texte et construit la séquence de phonèmes adéquate à prononcer avec une certaine prosodie. En général, ce bloc est décomposé en trois traitements :

1. le prétraitement qui remplace des éléments de texte particuliers, tels les acronymes, les nombres, les abréviations par un texte in extenso avant la lecture
2. la conversion de texte en phonèmes qui détermine les phonèmes à associer aux groupes de lettres
3. le générateur de prosodie en charge de l'intonation

Il existe différentes implémentations de ce traitement. En particulier, le package FreeTTS [5] qui est écrit en java et en libre distribution. Malheureusement, il ne prononce que l'anglais et n'utilise pas de règles de prononciation paramétrées. Il faut donc réécrire la plus grande part des classes java pour créer un TTS en français. De plus, l'architecture logicielle de FreeTTS est très complexe, ce qui a découragé nos étudiants. Une autre implémentation bien connue est fournie par le projet EULER [2]. Cependant, le logiciel est écrit en C++ ce qui n'est pas adapté à notre première année de cursus ingénieur.

Pour développer notre propre module de traduction en java nous avons décidé d'utiliser les règles de transcription en phonèmes de l'implémentation en PERL proposée par D. Haubensack [1]. Notre objectif étant de compléter cet ensemble de règles et d'améliorer la prosodie jugée insuffisante pour nos desseins.

3. Programmer la synthèse vocale

Nous présentons d'abord le convertisseur de texte en phonèmes puis la façon de générer la prosodie.

3.1 Conversion du texte en phonèmes

Règles de prononciations

Pour traduire un texte en une liste de phonèmes, au moins deux approches peuvent être choisies. Tout d'abord, donner un ensemble de règles de prononciations comme dans [6]. L'avantage est qu'il suffit de relativement peu de règles, mais l'inconvénient est que ces règles peuvent comporter des exceptions. La deuxième approche est de stocker la prononciation de tous les mots français (lemmes et règles d'inflexion) dans un dictionnaire [7]. Le problème est alors de se procurer un tel dictionnaire.

Nous avons décidé de fusionner ces deux approches : nous sommes partis de l'ensemble de règles de prononciation données dans le TTS en PERL [1] et l'avons enrichi en ajoutant un ensemble d'exceptions à ces règles. Règles de prononciation et exceptions sont décrites dans des fichiers de textes à l'aide d'une syntaxe simple et sont aisément modifiables par l'utilisateur qui peut ainsi améliorer la prononciation. Nous utilisons la syntaxe de [8] pour exprimer les règles et les exceptions:

prefixe [[*racine*]] *suffixe -> liste_phonème*

où *prefixe* et *suffixe* sont des expressions régulières. *Prefixe* est le contexte qui a été analysé avant la racine, *suffixe* est le contexte qui reste à analyser.

Par exemple, certaines règles de prononciation du groupe de lettres "am" sont présentées ci-dessous :

```
1   [[ am ]] n -> a m        ## amnistie
2   [[ am ]] m -> a          ## programmation
3   [[ am ]] C -> a~         ## camp
```

Dans ces règles, C est la classe des consonnes, ## est suivi d'un commentaire qui illustre la règle. Règle 1: *am* suivi de n est prononcé "a m". Règle 2: *am* suivi de m est prononcé "a" (on enlève simplement un m). Règle 3: *am* suivi d'une consonne est prononcé "an".

Création de la liste de phonèmes

Pour générer les phonèmes, nous cherchons dans l'ensemble des règles et des exceptions la règle dont la racine a le plus grand préfixe commun avec le mot à transcrire. Ensuite, nous vérifions si préfixes et suffixes concordent. Par exemple, pour générer les phonèmes du mot "camp", deux règles ont pour racine "c" qui est un préfixe de "camp" :

```
6   [[ c ]] (e|é|è|ê|i) → s      ## cesse
7   [[ c ]] → k
```

La règle 6 ne peut pas être appliquée car dans "camp" "c" est suivi de "a" (pas e ni é ni è …). Nous appliquons donc la règle 7 et générons le premier phonème **"k"**. Il faut ensuite analyser "amp". La plus longue règle qui s'applique est la règle 3 de l'exemple précédent et nous générons le phonème **"a~"**. Il faut maintenant traiter la dernière lettre "p" en appliquant la règle 8:

```
8   [[ p ]] T →
```

où T signifie "Terminal": un "p" à la fin d'un mot n'est pas prononcé. Nous obtenons enfin la liste: "**k a~"**.

Les règles de prononciation et leurs exceptions sont stockées dans la même structure de données et sont traitées en même temps. Voici par exemple des règles et exceptions pour les mots qui commencent par "qua":

 # règles générales

9 T [[quadr]] → k w a d R ## quadrature

10 T [[quadr]] il → k a d R ## quadrillage

11 [[qu]] → k

 # exceptions

12 T [[quantum]] T → k w a~ t o m

13 T [[quantifier]] T → k w a~ t i f i e

"quadr" est prononcé "k w a" (règle 9) excepté dans les mots de la famille de "quadrillage" (règle 10) comme "quadriller" ou "quadrille". Les autres mots comme "quand", "qualité", sont prononcés "k" avec la règle générale 11. Les règles 12 et 13 sont des exceptions pour les mots de la famille de "quantum" comme "quantifier" qui sont prononcés "k w a" (règles 12, 13).

Quand on analyse le mot "quantum" la règle 12 est utilisée (et pas la règle 11), car c'est celle dont la racine a le plus grand préfixe commun avec "quantum". Les règles de prononciation et les exceptions sont stockées dans des arbres lexicaux qui partagent les préfixes communs. Cette structure de données est peu coûteuse en complexité temporelle et spatiale. Les préfixes partagés sont les racines des règles et les noeuds terminaux contiennent la liste des règles associées à la racine (voir figure 1).

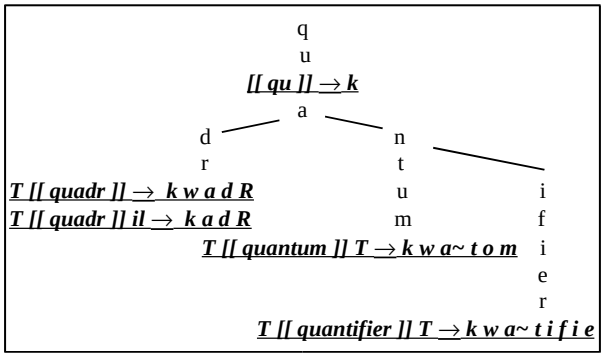


Figure 1: arbre lexical des règles 8, 9, 10, 11, et 12

En utilisant de tels arbres lexicaux, l'algorithme de transcription en phonèmes est tout simplement un parcours en profondeur de l'arbre, avec retour arrière si la règle ne peut pas s'appliquer à cause du suffixe. Les autres paramètres du fichier d'entrée de MBROLA sont générés par le module "prosodie".

3.2 Prosodie

La génération de la prosodie est un problème complexe. Notre objectif étant de concevoir un TTS utilisable dans des applications pour déficients visuels où les textes lus sont courts, nous avons adopté des principes simples. Nous avons affiné la micro prosodie qui est associée aux phonèmes dans l'implémentation en PERL. Par exemple, les phonèmes comme "a~" ou "o~" sont longs tandis que les phonèmes "R", "l", "H" sont courts. Le pitch des occlusives voisées, par exemple les lettres "b", "d" et "g" est choisi grave et la durée des derniers phonèmes des phrases est allongée.

Nous avons aussi ajouté une macro prosodie, grâce à une approche donnée dans [9]. Nous considérons cinq schémas intonatifs qui sont définis par des courbes qui s'échelonnent sur quatre niveaux de pitch. La distance entre ces niveaux ainsi que les pentes des courbes sont paramétrables. Pour appliquer ces schémas intonatifs, le texte est découpé en syntagmes qui sont identifiés soit par la ponctuation, soit par les conjonctions. Nous associons un schéma intonatif selon le type de syntagme (début, milieu ou fin de phrase). La durée de pause entre les syntagmes dépend de la ponctuation ou du type de conjonction. Par exemple, les conjonctions de subordination comme "bien que" sont traitées avec des pauses longues.

4. APPLICATIONS

Cette synthèse vocale est expérimentée depuis deux ans pour les applications de la journée DeViNT. D'autre part, elle est accessible sur http://kaivalyam.essi.fr (et référencée sur le site du projet MBROLA [1]) où elle a donné lieu à un ensemble de contacts, de suggestions et de retours de la part d'internautes et utilisateurs potentiels.

4.1. Les applications de la journée DeViNT

Les applications développées pour la journée DeViNT sont diverses, comme on le devine dans la liste non exhaustive présentée ci-dessous. Il s'agit souvent de jeux, destinés à des enfants de l'école primaire ou du collège. La synthèse vocale y joue un rôle fondamental pour l'accessibilité :

- CB2S : un TTS est associé à un lecteur de code barre. Ainsi les déficients visuels peuvent se faire lire toutes les informations sur les produits à partir d'un dispositif portable (tel PDA) qui interroge le serveur de synthèse vocale cité ci-dessus.
- Quinze questions : il s'agit d'un jeu de questions à choix multiples, inspiré du jeu télévisé « Qui veut gagner des millions ». Le TTS lit les questions et les réponses.
- Lecteur MP3 : l'objectif est de rendre accessible la musique MP3. Le TTS lit le titre des chansons.
- SI boud'chou : s'inspire du jeu ADIBOU pour apprendre le calcul et l'orthographe. Le TTS doit y lire des nombres (jeux de calcul) ou des mots (dictée).
- Installor : cet outil indispensable guide les déficients visuels lors de l'installation des logiciels DeViNT sur leurs propres ordinateurs.

4.2 Quelques retours d'expérience de DeViNT

Afin d'évaluer l'ergonomie et la pertinence de leurs logiciels, nos étudiants se sont rendus à l'institut Clément Ader à Nice présenter leurs projets en notant toutes les réactions des enfants. Ils ont ainsi récolté un ensemble d'informations pertinentes sur l'utilisation du son :

- il est souhaitable de développer une utilisation partagée de sons enregistrés et de la synthèse vocale. Les

enregistrements plus vivants sont utilisés pour les textes courts et répétitifs connus d'avance (par exemple, le texte des menus, les messages d'aide). Le TTS est utilisé pour les messages dépendants du contexte, ou qui sont inconnus a priori

- L'intonation monotone et machinale de la synthèse est jugée irritante à la longue, par les utilisateurs voyants. Mais il s'avère d'après certains utilisateurs non voyants que les voix plus performantes, telles 'Claire' ou 'Julie' ne sont pas toujours bien reçues pour d'autres raisons, en particulier le ton trop chantant qui est fatiguant. D'ailleurs, les aveugles choisissent une grande rapidité d'élocution pour briser la monotonie de ces voix synthétiques quand ils utilisent JAWS par exemple
- La clarté de la prononciation laisse parfois à désirer dans notre synthèse, sans pouvoir trancher s'il s'agit de la qualité de la base de diphones utilisée, ou la technique 'overlap and add' utilisée dans MBROLA pour accoler les diphones et synthétiser la voix
- Les règles de prononciation utilisées pour la traduction en phonèmes sont parfois insuffisantes. Il ne suffit pas toujours de compléter le fichier de règles : il faudrait dans certains cas utiliser un algorithme d'analyse syntaxique et pas seulement lexicale [3]

4.3 Les retours de : http://kaivalyam.essi.fr

Le TTS est accessible sur l'URL ci-dessus où il est possible de proposer un texte écrit et de récupérer le fichier audio correspondant. Le site permet aussi de contacter les auteurs pour faire des suggestions ou demander les sources java de la synthèse. Ce site a été maintes fois consulté, utilisé, et la synthèse demandée par des utilisateurs handicapés ou non, par des chercheurs, des internautes … et pas toujours pour résoudre la déficience visuelle. Citons par exemple un internaute privé temporairement de voix qui a utilisé notre synthèse pour communiquer par mail avec un enfant qui ne sait pas lire. Un autre exemple est l'utilisation dans le cadre du développement de sites citoyens pour le département du Val d'Oise, où une fonctionnalité de synthèse de la parole a été ajoutée pour donner un meilleur accès aux mal-voyants. Une autre demande a été faite par un élève de l'ENAC (Ecole Nationale de l'Aviation Civile à Toulouse) pour développer en java un simulateur de la gestion du traffic aérien, afin de préparer les étudiants aux épreuves pratiques.

D'autre part, le site a été utilisé pour faciliter le prototypage d'applications autour de DeViNT et de l'accessibilité. Dans ce cas les applications clientes se connectent au site en envoyant en paramètre le texte à lire et le serveur renvoie automatiquement le fichier wave contenant le texte lu. Par exemple, le lecteur de code barre CB2S, déjà cité et le projet EPUB de diffusion en temps réel des informations à l'institut Clément Ader ont utilisé cette fonctionnalité.

5. Conclusion et perspectives

La synthèse vocale que nous avons présentée répond à nos objectifs initiaux : elle est gratuite, portable, écrite en java, elle parle français. Elle a été intégrée dans plusieurs logiciels de jeux diffusés sur CDrom lors de la journée DeViNT et a donné satisfaction. Bien sûr la qualité de cette synthèse vocale pourrait être améliorée, en particulier pour la prosodie. N'étant pas spécialistes en la matière, nos efforts vont plutôt porter vers l'amélioration de l'ergonomie des applications dédiées aux déficients visuels. Tout d'abord en améliorant l'utilisation partagée des sons enregistrés et de synthèses vocales. Nous aimerions également réaliser une API permettant d'utiliser plusieurs synthèses vocales. L'idée est que nos étudiants puissent utiliser de façon transparente notre propre TTS (pour pouvoir le distribuer avec les jeux) mais également une ou des synthèses vocales commerciales de qualité qui seraient utilisées pour les démonstrations des journées DeViNT. Il nous faut aussi affiner l'analyse des retours d'expérience des utilisateurs déficients visuels.

Remerciements

Nous tenons à remercier les étudiants qui ont apporté leur enthousiasme et leurs compétences à ce projet, ainsi que nos collègues et en particulier M. Blay et A.M. Pinna qui sont à l'origine des journées DeViNT.

Bibliographie

[1] T. Dutoit & all, "The MBROLA project: Towards a Freely Available Multilingual Speech Synthesizer"
http://tcts.fpms.ac.be/synthesis/mbrola.html

[2] M. Bagein, T. Dutoit, F. Malfrère, A. Ruelle, N. Tounsi, D. Wynsberghe, *« The EULER Project, An Open, Generic, Multi-lingual and Multi-platform Text-To-Speech system", ProRISC'2000*

[3] J. P. Goldman, A. Gaudinat, L. Nerima, E. Wehrli, "FipsVox : a french TTS based on a syntactic parser", *4th ISCA int. Workshop on speech synthesis, 2001*

[4] T. Dutoit, *"An Introduction to Text-To-Speech Synthesis"*, Kluwer Academic Publishers, Dordrecht, 320 pp., ISBN 0-7923-4498-7, 1997

[5] W. Walker, P. Lamere, P. Kwok, "FreeTTS 1.2 - A speech synthesizer written entirely in the Java™ programming language", *Sun Mic. Lab. http://freetts.sourceforge.net*

[6] E. Keller, "Simplification of TTS architecture vs operational quality", *EUROSPEECH 1997*.

[7] F. Yvon, C. D'Alessandro, V. Aubergé, P. Boula de Mareüil, "Ressource standard pour le français : un large lexique orthographique-phonétique", *XXIIIèmes JEP, Aussois 2000*.

[8] N. Tounsi, R. Beaufort, « MLRR v1.1, The multi-Layer Rewriting rules parser », *TCTS, Faculté poly. de Mons http://tcts.fpms.ac.be/synthesis/mlrr/mlrr.html, Jul. 2001*.

[9] H. Fervers, J. Leroux, L. Miclet, « Programme de transcription phonémique en langue française », *Laboratoire de théorie des systèmes, Rapport ENST D-76003, 1977*

Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique ?

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

IRISA
Campus de Beaulieu, F-35042 Rennes Cedex, France
shuet@irisa.fr, ggravier@irisa.fr, sebillot@irisa.fr

ABSTRACT

The aim of the paper is to study the interest of part-of-speech (POS) tagging to improve speech recognition. We first evaluate the part of misrecognized words that can be corrected using POS information ; an analysis of a short extract proves that an absolute decrease of the word error rate by 1.1 % can be expected. We also demonstrate quantitatively that traditional POS taggers are reliable when applied to spoken corpus, including automatic transcriptions. This new result enables us to effectively use POS tag knowledge to improve, in a postprocessing stage, the quality of transcriptions, especially correcting agreement errors.

1. INTRODUCTION

Les systèmes de transcription utilisent globalement peu de connaissance sur le langage pour décoder la parole, se limitant en cela bien souvent au seul apprentissage des probabilités de successions de mots sur un corpus. Au vu du matériau manipulé, de la langue naturelle, il semble pourtant que des informations linguistiques supplémentaires devraient permettre d'améliorer la qualité de la transcription. Certains modèles de langage (ML) ont d'ailleurs déjà été conçus en intégrant des connaissances sur la structure syntaxique des groupes de souffle [2], sur les thèmes abordés par le document à transcrire [6] ou encore sur les parties du discours (appelées aussi *POS* pour *part of speech*) [7]. Une POS correspond à une propriété grammaticale d'un mot ou groupe de mots dans une phrase donnée (*e.g.* noms, verbes, prépositions, conjonctions, *etc.*), souvent accompagnée d'informations morphologiques (genre, nombre, conjugaison, *etc.*). La connaissance de ces catégories est généralement prise en compte au sein des ML à l'aide de modèles N-classes [1]. Si $\mathcal{C}_i$ représente l'ensemble des POS c_i auxquelles peut appartenir un mot w_i, le calcul des probabilités de la séquence de mots $w_1^n = w_1, ..., w_n$ s'effectue selon

$$P(w_1^n) \approx \sum_{c_1 \in \mathcal{C}_1, ..., c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i|c_{i-N+1}^{i-1}) \quad (1)$$

L'interpolation des modèles N-classes avec des modèles N-grammes conduit généralement à une baisse négligeable du taux d'erreur sur les mots de la transcription. Diverses améliorations ont donc été envisagées. Il a ainsi été proposé d'estimer la probabilité en considérant que les POS associées aux mots w_i à reconnaître font partie intégrante de la sortie de la transcription et ne sont plus un simple résultat intermédiaire [5]. Cette approche évalue les probabilités à l'aide d'un mode de calcul plus précis que celui de (1) mais conduit à une augmentation importante du nombre d'événements à considérer.

Dans ces diverses approches, les POS servent à construire des ML intervenant au cours du processus de transcription. Or ce type d'utilisation apporte un gain limité par rapport aux ML N-grammes de mots. Nous proposons donc, dans cet article, d'étudier la possibilité d'utiliser les étiquettes POS en aval de la transcription, pour sélectionner la meilleure hypothèse parmi plusieurs proposées par le système de reconnaissance. La première étape de notre travail a consisté à déterminer la proportion des erreurs de transcription corrigeable par la connaissance des POS. Celle-ci étant importante, nous avons cherché à évaluer la capacité des méthodes automatiques à étiqueter des transcriptions. Les étiqueteurs sont en effet conçus à l'origine pour des documents écrits, dont certaines caractéristiques sont très différentes d'un corpus oral, d'autant plus si celui-ci comporte des erreurs de transcription. Nos différentes évaluations ayant montré l'aptitude des étiqueteurs, nous avons mené de premières expérimentations pour tester l'utilisation des POS en post-traitement du système de transcription. Le plan de la suite de l'article suit les différentes étapes de notre démarche.

2. TYPOLOGIE DES ERREURS DE TRANSCRIPTION

Afin d'évaluer l'apport potentiel des POS pour la transcription, nous avons étudié en détail un court extrait de transcription automatique, en cherchant à connaître la part des erreurs corrigeables par cette seule connaissance.

Le système de reconnaissance utilisé dans nos expérimentations, développé par l'IRISA et l'ENST pour la campagne ESTER, permet de produire un graphe de mots en trois passes[1]. Un premier graphe de mots est généré avec des modèles acoustiques hors-contexte et un ML trigramme. La deuxième passe utilise des modèles contextuels pour réévaluer les 1 000 meilleurs chemins extraits du graphe de la première passe à l'aide d'un ML 4-gramme. Enfin, la troisième passe, similaire à la deuxième après adaptation au locuteur des modèles contextuels, permet de générer un graphe d'hypothèses. Afin de s'affranchir des problèmes de segmentation, nous considérons dans ce travail une segmentation manuelle en groupes de souffle basée sur la détection de pauses silencieuses. Nous examinons ici un extrait de 6 500 mots d'une émission d'information sur France-Inter, issue du corpus ES-

[1]Nous tenons à remercier François Yvon pour nous avoir fourni le ML et le lexique étiqueté.

TER [4]. Le taux d'erreur sur cet extrait est de 17,8%.

Parmi les erreurs de reconnaissance que nous y avons constatées, trois groupes se détachent. Certaines erreurs correspondent à un « dérapage » du système, généralement dû soit à une mauvaise acoustique, soit à une mauvaise reconnaissance d'entités nommées. Ces erreurs semblent hors d'atteinte de la correction susceptible d'être apportée par les POS. Heureusement, elles ne concernent qu'une part très restreinte de l'extrait analysé. Le deuxième ensemble correspond à des groupes de souffle agrammaticaux (Fig. 1). L'agrammaticalité est notamment causée par les mots grammaticaux « *a* », « *à* », « *de* », « *que* » ou encore « *et* », qui sont parfois absents, ou présents de manière inopinée dans les hypothèses de transcription. On retrouve également des fautes de temps et de mode des verbes, le présent et l'indicatif étant souvent privilégiés. Parmi ces erreurs, certaines semblent corrigeables puisque l'étiquetage des groupes de souffle peut conduire à des séquences de POS aberrantes, comme l'apparition de trois prépositions consécutives. Ce critère est néanmoins à prendre avec précaution, à cause des répétitions présentes dans la langue parlée. Le troisième groupe est formé d'erreurs très vraisemblablement corrigeables grâce aux POS, à savoir les fautes d'accord en genre et en nombre et les confusions entre infinitif et participe passé. Ces erreurs sont particulièrement nombreuses puisqu'elles concernent un groupe de souffle sur sept. Parmi elles, 70 sont rectifiables sans avoir à examiner de dépendances entre des groupes de souffle consécutifs (Fig. 1), et les corriger représenterait une baisse absolue de 1,1% du taux d'erreur. Au travers de l'exposé des principales erreurs de décodage, il apparaît donc que les POS constituent une source d'information intéressante pour améliorer la qualité de la transcription.

3. COMPORTEMENT DES ÉTIQUETEURS

La section précédente a montré l'intérêt des étiquettes POS pour corriger des erreurs de transcription en se focalisant sur des successions de POS possibles. Pour pouvoir utiliser cette technique, il faut cependant que les étiqueteurs fonctionnent de manière fiable sur des corpus oraux produits par des annotateurs ou obtenus par des systèmes de reconnaissance. C'est cette propriété que nous cherchons à évaluer ici.

Le rôle des étiqueteurs morphosyntaxiques est d'associer à chaque mot ou groupe de mots de la séquence à étudier l'étiquette catégorielle la plus probable. Ces outils sont ordinairement appliqués sur des corpus écrits. Pour les évaluer quantitativement, un texte est étiqueté manuellement par des annotateurs et les étiquettes sont comparées une à une avec celles proposées par la méthode automatique. Comparativement aux corpus écrits, les corpus oraux, transcrits par des annotateurs, ont été peu étudiés [8]. La production orale présente des caractéristiques, telles que les reprises ou les répétitions, qui sont susceptibles de compliquer l'opération d'étiquetage. L'étiquetage de la transcription automatique de la parole planifiée est une tâche rendue plus complexe encore par le fait que le texte est segmenté en groupes de souffle et non en phrases, et ne contient ni ponctuations, ni majuscules (cas du vocabulaire de notre système de transcription).

De manière à faciliter l'utilisation des POS pour décoder la parole, nous avons fait le choix de construire notre propre étiqueteur morphosyntaxique. La suite de cette section décrit le protocole utilisé pour construire cet étiqueteur, avant d'évaluer son comportement sur de la parole transcrite. Nous avons examiné la qualité de l'étiquetage produit sur un corpus de test et l'avons comparée avec les résultats obtenus avec un étiqueteur qui fait référence pour le français.

3.1. Constitution d'un étiqueteur

Les étiqueteurs conçus pour l'écrit utilisent des règles linguistiques ou extraient automatiquement l'information statistique contenue dans de grands volumes de données. Dans la mesure où les programmes basés sur des calculs statistiques conduisent à des résultats satisfaisants pour l'écrit et ne nécessitent pas l'écriture manuelle de nombreuses règles contextuelles, nous avons construit notre étiqueteur en utilisant exclusivement des méthodes statistiques.

Pour ce faire, nous avons constitué un corpus d'apprentissage de 200 000 mots représentant un extrait d'une durée de 16 heures du corpus ESTER. Les transcriptions manuelles, contenant à l'origine des majuscules et des ponctuations, ont été étiquetées par le logiciel Cordial[2]. Le résultat a été vérifié manuellement, puis nous avons ôté toutes les majuscules et les marques de ponctuation pour qu'il soit cohérent avec la forme du texte produit par notre système de reconnaissance. Nous avons utilisé un lexique de prononciations étiqueté afin de connaître les POS possibles pour chaque mot. Le choix des étiquettes morphosyntaxiques a été fait de manière à discriminer le genre et le nombre des adjectifs et des noms, et le temps et le mode des verbes, ce qui conduit à un jeu de 80 étiquettes différentes. Cet ensemble d'étiquettes est très proche de celui proposé dans les grammaires scolaires et est directement inspiré de celui utilisé par Cordial.

Notre étiqueteur morphosyntaxique se base sur un modèle N-classe pour trouver la séquence d'étiquettes qui maximise le produit dans (1) pour une séquence de mot w_1^n. Des réglages sur un corpus de développement nous ont conduit à choisir un ordre $N = 7$ et un lissage de Kneser-Ney non modifié. De façon à évaluer l'impact de la segmentation sur la qualité de l'étiquetage, nous avons procédé à deux apprentissages différents, en segmentant le corpus d'apprentissage par phrases puis par groupes de souffle.

3.2. Évaluation de l'étiquetage

Afin d'avoir une mesure quantitative de la qualité de l'étiquetage sur des transcriptions produites manuellement (REF), segmentées en phrases ou en groupes de souffle, ou produites automatiquement (HYP) par le système de reconnaissance, nous avons étiqueté manuellement une émission d'information de France-Inter d'une heure, constituée de 11 300 mots, que nous désignerons par GOLD. L'étiquetage automatique de REF a été évalué en dénombrant le nombre d'étiquettes en commun avec GOLD. La mesure de la qualité de l'étiquetage a été plus problématique pour HYP, pour lequel nous avons mesuré un taux d'erreur de transcription de 22%, puisque les mots ne sont pas identiques avec ceux du GOLD. Il a ainsi été impossible de constituer un étiquetage de référence pour HYP dans la mesure où il n'existe pas de POS cohérentes pour les mots des groupes de souffle agrammaticaux. Nous

[2] Version 8.1 distribuée par la société *Synapse Développement*.

Hypothèse agrammaticale
REF: bush ** SAIT donc QU' il faudra coopérer
HYP: bush s' EST donc ** il faudra coopérer
Erreur d'accord
REF: c' est un monstre injuste envers sa soeur si DÉVOUÉE
HYP: c' est un monstre injuste envers sa soeur si DÉVOUÉ

FIG. 1: Exemples d'erreurs dans les groupes de souffle

donnons donc deux mesures de la qualité de l'étiquetage de HYP : le pourcentage de mots correctement reconnus et étiquetés parmi le nombre total de mots du GOLD, et le pourcentage de mots correctement reconnus et étiquetés parmi le nombre de mots bien reconnus dans HYP (donné entre parenthèses dans le tableau 1).

Les résultats obtenus par notre étiqueteur sur les corpus de test sont présentés dans les deux premières lignes du tableau 1, en effectuant toutes les compositions possibles en ce qui concerne la segmentation du corpus d'apprentissage et des corpus de test. Ils établissent que l'étiquetage produit est bon dans l'ensemble, y compris pour les transcriptions automatiques dont les erreurs de reconnaissance peuvent venir perturber l'étiquetage des mots correctement transcrits. Ces résultats sont relativement surprenants dans la mesure où nous n'avons pas introduit de méthodes spécifiques pour traiter les particularités de la langue orale, si ce n'est d'utiliser un corpus oral pour paramétrer l'étiqueteur. Cette robustesse des étiqueteurs sur la langue parlée s'explique cependant par le fait que les étiquettes sont attribuées en exploitant des informations de manière locale. Il apparaît en outre que l'apprentissage à partir d'une segmentation en groupes de souffle fournit les meilleurs résultats, ce qui nous a conduit à privilégier ce mode de segmentation par la suite.

De plus, en examinant les fautes commises dans l'attribution des POS, nous avons constaté que certaines pouvaient être considérées comme acceptables. Ainsi, les distinctions entre les POS « participe passé » et « adjectif » sont dans la grande majorité des cas discutables. Nous avons également constaté de nombreuses erreurs dues à la mauvaise tokenisation de notre étiqueteur. Ainsi, alors que le GOLD avait étiqueté respectivement « *états-unis* » et « *alors que* » comme nom propre et conjonction de subordination, l'étiquetage automatique a conduit à reconnaître d'une part « *états* » comme nom commun, « *unis* » comme adjectif et, d'autre part, « *alors* » comme adverbe et « *que* » comme conjonction de subordination. Sur les 966 erreurs observées lors de l'étiquetage de REF segmenté par groupes de souffle, 42 sont dues à des confusions entre participe passé et adjectif, 216 à des erreurs de tokenisation, 124 à des confusions entre nom commun et nom propre et 10 à des mots inconnus par l'étiqueteur.

Nous avons en outre comparé les performances de notre étiqueteur à celles de Cordial, vraisemblablement meilleur étiqueteur disponible pour le français écrit, qui a déjà donné de bons résultats sur un corpus de parole [8]. La dernière ligne du tableau 1 présente ses résultats sur le corpus de test. L'examen de ce tableau établit que notre étiqueteur a des résultats comparables, voire meilleurs que Cordial. On peut d'ailleurs constater que Cordial se comporte moins bien qu'habituellement, ses scores sur de l'écrit étant généralement supérieurs à 95%. Ceci s'explique par la nature particulière de la transcription automatique, pour laquelle il n'a pas été spécifiquement conçu. L'absence de majuscules est particulièrement problématique dans la mesure où le logiciel s'appuie sur cet indice pour détecter les noms propres. En ignorant toutes les erreurs dues à une confusion entre nom commun et nom propre, le pourcentage d'étiquettes bien attribuées monte à 93,52% pour le corpus de test segmenté par groupes de souffle, alors que, suivant le même critère et sur les mêmes données de test, les performances de notre étiqueteur ne progressent qu'à 92,55%.

Cette série d'expérimentations montre que l'étiquetage des transcriptions automatiques est fiable, ce qui n'était encore qu'une hypothèse auparavant. Notre étiqueteur conduit à des résultats qui nous permettent de l'envisager pour calculer un score sur la qualité du décodage. La section suivante présente des résultats préliminaires sur l'utilisation de la connaissance des POS en post-traitement de la transcription.

4. APPORT DE L'ÉTIQUETAGE À LA TRANSCRIPTION

De manière à exploiter la connaissance des POS au cours du décodage de la parole, nous avons employé notre étiqueteur pour attribuer un score à chaque hypothèse donnée sur un groupe de souffle. Chaque hypothèse candidate w_1^n est étiquetée par la séquence de POS c_1^n la plus probable, avant d'être évaluée par la fonction de score suivante :

$$lp = \log P(c_1^n) = \sum_{i=1}^{n} \log P(c_i | c_{i-N+1}^{i-1}) \qquad (2)$$

Ce score vise à réévaluer la liste des meilleures hypothèses produites par le système de reconnaissance pour chaque groupe de souffle.

Afin de valider notre approche, nous avons testé dans un premier temps le comportement sur les 70 erreurs d'accord que nous avions repérées (*cf.* section 2). Pour chacun des groupes de souffle contenant l'une de ces erreurs, nous avons établi le score pour trois versions : la transcription de référence (REF), la transcription automatique (HYP) et la transcription automatique où seules les erreurs d'accord sont corrigées (COR) (Fig. 2). Nous espérons ainsi que la succession d'étiquettes obtenues sur REF et COR sera plus probable que sur HYP.

Nous avons constaté sur les 63 groupes de souffle analysés que le score était meilleur sur COR que sur HYP pour 46 d'entre eux et meilleur sur REF que sur HYP pour 41 d'entre eux. Ces résultats établissent ainsi que, dans une majorité des cas, lp conduit à une correction des fautes d'accord et est donc susceptible d'apporter un gain au niveau du taux d'erreur de reconnaissance.

TAB. 1: Évaluation des étiqueteurs (en pourcentages)

	REF / phrase	REF / groupe de souffle	HYP
corpus d'app / phrase	91,42	91,09	72,60 (91,83)
corpus d'app / groupe de souffle	91,50	91,42	72,99 (92,32)
Cordial	88.69	88,61	70,75 (89,48)

```
REF: à  L' AMÉNAGER avant qu' elle ne soit DÉTRUITE
COR: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUITE
HYP: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUIT
```

FIG. 2: Versions à évaluer pour un même groupe de souffle

Nous avons utilisé ce score pour réordonner la liste des 100 meilleures hypothèses produites pour 4 heures d'émissions d'informations en français. Le taux d'erreur sur les mots donné par un oracle sur cette liste est de 14,2 % pour un taux initial de 21,6 %. En réordonnant la liste en utilisant lp, le taux augmente de manière significative de 21,6 % à 26,2 %. Nous avons donc décidé de combiner le score sur les POS avec le score acoustique et le score du ML.

La reconnaissance de la parole est en pratique habituellement exprimée comme une recherche de w_1^n à partir de l'entrée acoustique y_1^n. Pour introduire lp, nous modifions le critère de sélection de w_1^n par

$$\hat{w}_1^n = \arg\max_{w_i^n} [\log P(y_1^n|w_1^n) + \alpha \log P(w_1^n) + \beta \log P(c_1^n) + \gamma n] \quad (3)$$

où α est le facteur d'échelle du ML et γ est la pénalité d'insertion d'un mot. $P(w_1^n)$ est calculée par un ML 4-gramme sur les mots, tandis que $P(c_1^n)$ est déterminée par un ML 7-gramme sur les POS.

Nous avons observé une légère diminution du taux d'erreur à 21,4 % avec cette méthode. Nous avons également remarqué que généralement les erreurs d'accord ont été corrigées. Par exemple, le groupe de souffle initialement transcrit par « *le messin disputent aujourd'hui* » a été correctement rectifié par « *le messin dispute aujourd'hui* ». Toutefois, quelques erreurs apparaissent comme la transcription, correcte au départ, « *les visages de Jacques Chirac et Jean-Marie Le Pen apparaissent* » qui a été modifiée en « *les visages de Jacques Chirac et Jean-Marie Le Pen apparaît* ».

La connaissance des POS apporte donc une information réduite par rapport au ML basé sur les mots, bien que les deux méthodes soient complémentaires comme le montre la réduction du taux d'erreur. Notre approche semble en outre un peu plus performante que celle des ML N-classes classiques puisqu'en effectuant une combinaison linéaire de ce type de modèle avec un ML 4-gramme, et ce, en utilisant le même jeu d'étiquettes que précédemment, le taux d'erreur n'a pu être réduit en dessous de 21,6 %.

5. PERSPECTIVES

Dans cet article, nous avons d'une part montré l'intérêt de la connaissance des étiquettes POS pour corriger des erreurs de transcription de parole pour le français et avons, d'autre part, prouvé quantitativement que les étiqueteurs pouvaient réellement être utilisés sur des corpus oraux transcrits manuellement ou automatiquement, ce qui rend effectivement possible l'exploitation des POS pour améliorer les transcriptions. De premières expériences ont montré que les étiquettes POS pouvaient corriger des fautes d'accord, même si cela se manifeste globalement par une diminution modeste du taux d'erreur sur les mots. Pour améliorer ces premiers résultats, au lieu d'opérer sur les N meilleures hypothèses du système de transcription, nous prévoyons de réévaluer tous les homophones de la meilleure hypothèse trouvée [3]. En outre, nous souhaitons étudier l'influence d'autres jeux d'étiquettes POS sur la qualité de la transcription.

RÉFÉRENCES

[1] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992.

[2] C. Chelba and F. Jelinek. Structured language modeling. *Computer Speech and Language*, 14(4):283–332, 2000.

[3] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. Where are we in transcribing French broadcast news? In *Proc. of Eurospeech*, 2005.

[4] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. In *Actes des JEP*, 2004.

[5] P.A. Heeman. POS tags and decision trees for language modeling. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[6] S. Khudanpur and J. Wu. A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition. In *Proc. of ICASSP*, 1999.

[7] G. Maltese and F. Mancini. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In *Proc. of ICASSP*, volume 1, 1992.

[8] A. Valli and J. Véronis. Étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133, 1999.

Algorithme de recherche d'un rang de prédiction. Application à l'évaluation de modèles de langage

P. Alain, O. Boëffard

IRISA / Université de Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
{pierre.alain,olivier.boeffard}@irisa.fr
http ://www.irisa.fr/cordial/

ABSTRACT

Within a predictive framework for language model evaluation, Shannon uses a rank distribution in order to bound the entropy of printed English. Taking into account of higher dimensions (prediction of symbols in a raw) and predicting a k-word sequence given a N-word vocabulary is a NP-hard computational task. To achieve this goal, we propose some acceptable and effective search heuristics for an A^* algorithm.

1. INTRODUCTION

De nombreux systèmes de traitement de l'information font appel aux modèles de langage. Les domaines d'application sont variés, il s'agit notamment du traitement de la parole, de la traduction automatique, ou de la recherche d'information. L'évaluation des modèles de langage repose sur deux postures méthodologiques relativement tranchées. Une première adopte pour principe qu'il est nécessaire d'évaluer un modèle dans son contexte d'application. Rosenfeld [7] considère à cet égard qu'il faut parler d'un gain de 10% sur la perplexité pour obtenir des résultats significatifs notamment en reconnaissance de la parole. Une seconde approche définit un cadre méthodologique qui ne nécessite pas l'application visée. Le critère le communément admis est celui de l'entropie-croisée, exprimé sous la forme d'une mesure de perplexité. Le principal inconvénient du critère de perplexité est qu'il fait l'hypothèse d'un modèle probabiliste. Comme corollaire à cette nécessité, la mesure de perplexité est directement reliée au nombre de paramètres du modèles. Notamment, plus le nombre de degré de liberté d'un modèle augmente meilleurs sont les scores de perplexité. Avec ce seul critère, il est donc impossible de juger à la fois de la précision d'un modèle et de son efficacité (au sens d'une longueur de description minimale par exemple), sauf à mettre en place un scénario d'évaluation plus complexe de validation croisée.

Les travaux que nous proposons se situent sur le deuxième axe méthodologique. Nous pensons en effet qu'il est plus pertinent d'évaluer un modèle pour ses qualités propres qu'au travers d'une application. Une évaluation qui dépend de la tâche ne permet pas, en toute rigueur, de généraliser des résultats à des domaines connexes. Dans [3], nous avions fait le choix de juger de la performance d'un modèle de langage selon ses capacités de prédiction. Ce scénario méthodologique reprend les idées de Shannon [8], nous les avons étendues à la prédiction conjointe de séquences de mots de manière à couvrir les différents types de modèle de langage proposés par la littérature (notamment ceux qui prédisent plusieurs mots en un seule étape de traitement). Bimbot [2] a déjà fait allusion au cadre de Shannon mais reste sur une estimation d'une mesure de perplexité.

Cet article propose une solution algorithmique à l'évaluation du rang de prédiction d'une séquence de mots étant connu un historique. Le prédicteur, ou modèle de langage, est ici très général, il peut s'agir de modèles probabilistes ou non. Le problème est de nature combinatoire. La difficulté provient du fait que la décision de prédiction sur un mot dépend des prédictions déjà réalisées. Nous n'avons pas la place de présenter à la fois les détails algorithmiques et une évaluation expérimentale. Nous présentons ici les démonstrations d'admissibilité des fonctions d'élagage appliquées à l'algorithme de recherche des meilleurs chemins construit sur une base A^*.

Au cours du paragraphe 2, nous présentons la problématique du calcul des rangs de prédiction ainsi qu'une formalisation par des graphes multivalués. Dans la section 3 nous présentons des algorithmes de parcours pour un graphe mulivalué. La section 4 présente les fonctions d'élagage qui ont été mises en place. Nous conclurons dans la section 5.

2. MODÉLISATION DU PROBLÈME

Soit une séquence de l mots de test, la détermination du rang de prédiction de cette séquence par un modèle de langage passe par une modélisation sous forme de graphe. On reprend les notations proposées par [1]. On note G un graphe, s le nœud source qui relie la fin des mots de l'historique au début de la fenêtre de prédiction. On note $c(i,j)$ le coût d'un arc du nœud i au nœud j, il s'agit donc du coût d'émission d'un mot, et $c(P,i)$ le coût entre s et i sur le chemin P.

La figure 1, partie $a)$ présente le graphe qui correspond à la recherche des chemins de prédiction des trois mots $[W_i\ W_{i+1}\ W_{i+2}]$. Cette prédiction peut s'effectuer mot-à-mot en utilisant les arcs de coût $b(W_i)$, $b(W_{i+1})$ ou encore $b(W_{i+2})$. Ce type de cheminement dans le graphe est isomorphe à un calcul de perplexité. On peut également choisir de prédire deux mots en une seule fois avec les arcs de coût $b(W_i, W_{i+1})$ ou $b(W_{i+1}, W_{i+2})$. Enfin, on peut directement prédire les trois mots en une seule fois avec l'arc de coût $b(W_i, W_{i+1}, W_{i+2})$.

Ce graphe est *multivalué*. Chaque arc représente en réalité un fuseau d'arcs qui correspond à l'ensemble des historiques possibles que le modèle de langage peut utiliser pour effectuer sa prédiction, figure 1 partie $b)$. Les arcs

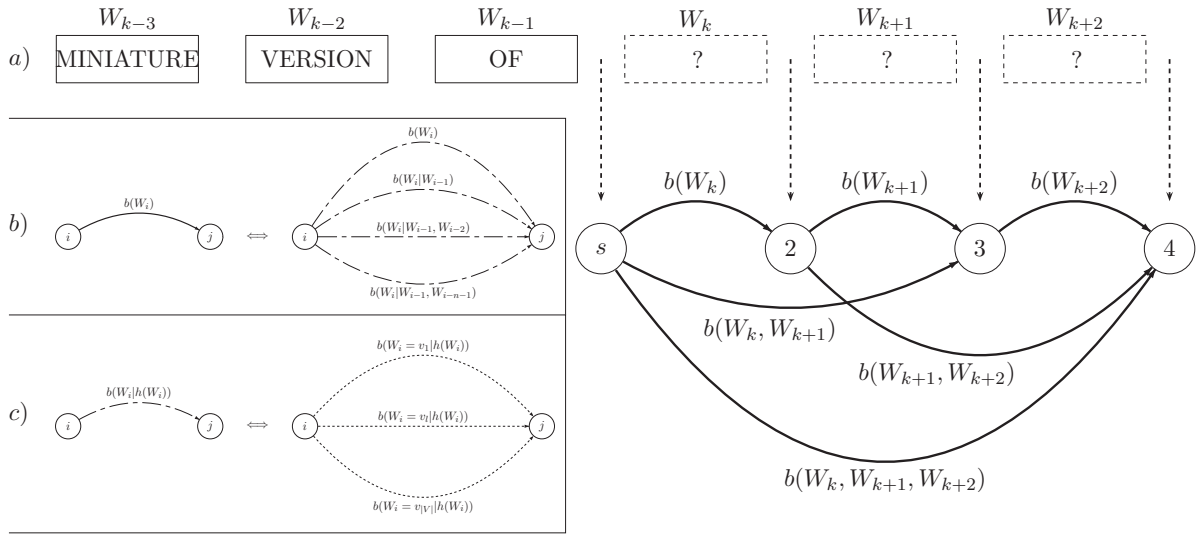


FIG. 1: Le graphe de recherche de chemins pour la prédiction de trois mots à partir d'un contexte connu $[MINIATURE\ VERSION\ OF]$ (a)). La partie $b)$ présente le fuseau en fonction de l'historique, la partie $c)$ le fuseau en fonction du mot prédit.

qui composent ces fuseaux sont eux-mêmes des fuseaux d'arcs. Chaque arc, au dernier niveau, correspond à l'ensemble des mots connus du modèle ; c'est-à-dire à l'ensemble des mots du vocabulaire, plus précisément à l'ensemble des mots prédictibles, figure 1 partie $c)$.
Le graphe de prédiction est un graphe multivalué *dirigé*, la prédiction se fait toujours sur un ou plusieurs mots du futur, les arcs sont alors tous dirigés dans le même sens. Le graphe est également *non absorbant*, car le coût d'un arc est toujours de même signe. En empruntant un nouvel arc, on ne baisse jamais le coût total du chemin.
La propriété de multivaluation entraîne un *non déterminisme*. Il existe plusieurs arcs qui permettent de prédire le mot W_i. Ce graphe peut être transformé en un graphe déterministe par démultiplication des nœuds en fonction de l'historique et du mot prédit. Le graphe déterministe ainsi construit possède un nombre exponentiel de nœuds.
Ce graphe est également *dynamique*, en effet, un modèle de langage tenant compte d'un historique peut donner un coût $c(i, j)$ différent suivant le chemin emprunté pour arriver au nœud i.

La détermination du rang de prédiction R d'une séquence de mots de test correspond à la recherche de tous les meilleurs chemins de prédiction du modèle de langage jusqu'au chemin qui correspond à l'émission des mots de test. Le rang de prédiction R est donné par le rang de classement du chemin des mots de test. Le rang 1 correspond à l'émission la plus probable du modèle de langage (qui n'est pas forcément la séquence des mots de test).

Nous venons de caractériser par un graphe un modèle de description qui tient compte des différentes alternatives proposées par un modèles de langage. Il nous reste à définir une procédure d'énumération de ces alternatives de prédiction pour ensuite les classer entre elles en fonction de leurs rangs.

3. PARCOURS DU GRAPHE

On s'intéresse tout d'abord aux algorithmes de recherche de la meilleure séquence de prédiction. La recherche de cette meilleure séquence correspond à une recherche du meilleur chemin dans le graphe décrit précédemment. On peut montrer que des algorithmes de décodage classiques, comme l'algorithme de Dijkstra, ne peuvent cependant pas s'appliquer à ce graphe dynamique.

Nous avons dans un premier temps adapté un algorithme A^* pour la recherche d'un meilleur chemin dans un graphe multivalué, section 3.1. Cette première base algorithmique a ensuite été complétée pour le traitement des R meilleurs chemins, section 3.2. Cette section montre comment cet algorithme ne peut éviter l'explosion combinatoire lors de la recherche car nous ne connaissons pas la valeur de coût du chemin au rang R. L'alternative que nous proposons à cette contrainte combinatoire sera développée section 4 et consiste en une fonction efficace d'élagage des chemins de rang supérieur à celui de la séquence des mots de test.

3.1. Algorithme A^*, principe

L'algorithme A^*, dans le cas général, conserve au niveau de chaque nœud un ensemble d'hypothèses. Cet ensemble est une mémoire des meilleurs chemins qui permettent d'atteindre un nœud n depuis le nœud s. Deux listes globales permettent de conserver ces informations : une liste $OPEN$ de nœuds à explorer et une liste $CLOSE$ de noeuds déjà explorés.

A^* se base sur une heuristique qui estime le coût du chemin optimal d'un nœud n au nœud final. Pour donner le résultat optimal, cette heuristique ne doit en aucun cas surestimer le coût global jusqu'au nœud final. L'algorithme est alors dit admissible. Si $h^*(n)$ est le coût réel du chemin du mot n jusqu'à la fin de la fenêtre de prédiction, et $h(n)$ le coût proposé par une heuristique admissible : $h(n) \leq h^*(n)$) [4].

3.2. Recherche des R meilleurs chemins

L'algorithme A^* garde sur chaque nœud une liste ordonnée des meilleurs hypothèses qui permettent d'atteindre un nœud n depuis le nœud source. Il est donc très facile de passer d'une version algorithmique de recherche d'un meilleur chemin à une recherche des R meilleurs [5, 6].

Le déroulement de l'algorithme est le suivant : on initialise un tas avec le nœud s de coût d'accès nul, puis on itère sur le procédé suivant : on enlève du tas le chemin de coût le plus faible (en fait le nœud final du chemin dont le coût d'accès est le plus faible), puis on met dans le tas tous les candidats, c'est à dire tous mots possibles selon le modèle, qui peuvent apparaître à la suite du nœud extrait. Une heuristique permet de tenir compte du coût reel du chemin déjà parcouru, $c(P, i)$, du coût de au nœud voisin ($c(i, j)$), et d'une estimation du coût du chemin jusqu'au nœud de destination ($h(j)$) . Comme déjà discuté, on prendra la valeur nulle comme estimation du coût futur. On obtient ainsi une implantation capable de fonctionner avec un graphe multivalué.

À l'exécution, ajouter tous les candidats possibles fait croître très rapidement la taille du tas. Pour cette raison, un élagage doit être entrepris. Le principe consiste à vérifier que le coût d'un chemin passant par le nœud courant jusqu'au nœud voisin reste inférieur à une certaine valeur de seuil. Toute la difficulté de cette méthodologie est que lors de la recherche des meilleurs chemins, on ne connaît pas encore R. Si nous ne recherchons que les k meilleurs chemins, nous devons nous assurer que le rang de la fenêtre de prediction R est bien tel que $k \geq R$.

4. A^* ET HEURISTIQUE D'ÉLAGAGE

Nous avons constaté qu'un graphe multivalué peut conduire à plusieurs chemins possibles pour une séquence de mots de test. Pour une fenêtre de prédiction, le rang des mots de la séquence de test, $W_i, \cdots, W_{i+l}$, est celui du chemin qui émet cette séquence avec le coût le plus faible. Nous présentons tout d'abord, section 4.1, l'algorithme A^* mis en œuvre. Ensuite deux heuristiques d'élagage sont proposées pour limiter l'explosion combinatoire de la recherche de R meilleurs chemins, section 4.2.

4.1. Présentation de l'algorithme

Le principe de l'élagage consiste à ne pas inclure dans la liste $OPEN$ des nœuds qui appartiennent à des chemins dont le rang sera supérieur au coût le plus faible des chemins qui correspondent aux mots de test de la fenêtre de prediction. C'est-à-dire qu'aucun chemin en cours de construction par A^* ne doit être éliminé s'il peut obtenir un rang meilleur que celui de la fenêtre de prediction. L'algorithme utilisé est l'algorithme 1.

N_{sol} représente le nombre de solutions déjà trouvées, Sol est l'ensemble de ces solutions. $ELAGAGE$ correspond à l'heuristique d'élagage, voir section 4.2 pour les détails d'implantation. $heap \Leftarrow \{s\}$ indique que le tas est initialisé avec le nœud source, noeud fictif qui correspond à la fin du contexte connu (nœud 1 sur la figure 1). Parmi les opérateurs utilisés, $last(Sol)$ donne la dernière solution ajoutée dans Sol si Sol est non vide, $null$ sinon. $is\_sol(.)$ renvoie vrai si le paramètre en entrée correspond à la sequence de prédiction ($is\_sol(null) = false$). $min(.)$ retourne et supprime du tas le chemin de coût minimal. $length(.)$ donne la longueur du chemin passé en paramètre. $next(.)$ est une fonction qui à partir d'un nœud donne la liste de ses voisins : cette fonction permet de traverser le graphe. $\oplus$ permet d'ajouter un nœud à un chemin.

Algorithme 1

```
1   N_sol ⇐ 0
2   Sol ⇐ ∅
3   OPEN ⇐ {s}
4   while (¬(heap = ∅ ∨ is_sol(last(Sol)) ∨
5                          N_sol = N_max)) do begin
6    path ⇐ min(OPEN)
7    if (length(path) = l) begin
8     Sol ⇐ Sol ∪ {path}
9     N_sol ⇐ N_sol + 1
10   end else begin
11    forall (node ∈ next(path)) do begin
12     P ⇐ path ⊕ node
13     if (c(P, node) < ELAGAGE) begin
14      OPEN ⇐ OPEN ∪ {P}
15     end
16    end
17   end
18  end

19  if (is_sol(last(Sol))) begin
20   Le rang de la séquence est N_sol
21  end else begin
22   La séquence n'a pas été trouvée
23  end
```

La paragraphe suivant établit des propositions pour donner une valeur au seuil d'élagage $ELAGAGE$.

4.2. Fonctions d'élagage

Avant de décrire et comparer les deux fonctions d'élagage proposées, la proposition suivante définit un domaine de recherche de ces constantes.

Proposition 1 *Soit une fonction d'élagage constante c supérieure au coût du chemin des mots de test dans la fenêtre de prédiction. Alors les seuls chemins élagués sont de rangs strictement supérieurs à celui de la fenêtre de prédiction.*

Preuve : Soit P un chemin élagué par la fonction d'élagage c, le coût de P est donc strictement supérieur à c. D'après le choix de c, le chemin P a un coût strictement supérieur à celui de la fenêtre de prédiction. Par conséquent, le rang de P est strictement supérieur à celui de la fenêtre de prédiction. □

On peut donc choisir comme premier seuil d'élagage le coût du chemin qui correspond à la prédiction des mots de test. On se retrouve à poser comme heuristique le coût d'une séquence de mots évalué mot à mot.

En partant du premier nœud, jonction entre l'historique connu et le début de la fenêtre de prédiction, on demande au modèle de langage d'évaluer le coût de prédiction du premier mot de la fenêtre. On ajoute ce premier mot à l'historique dont se sert le modèle pour évaluer le coût de prédiction du mot suivant. On continue ainsi jusqu'à la fin de la fenêtre. On obtient alors une évaluation du coût c_1 de prédiction de cette fenêtre de test. On définit ainsi la

fonction d'élagage constante c_1. On rappelle que le coût de la fenêtre de prédiction est défini comme étant le minimum des coûts des chemins menant à l'émission des mots de test de la fenêtre de prédiction. Par conséquent, la fonction d'élagage c_1 vérifie la proposition 1.

Une seconde proposition consiste à exécuter l'algorithme de Dijkstra de manière à ce que le coût de prédiction intègre toutes les alternatives possibles du modèle de langage, notamment la prédiction conjointe de plusieurs mots. L'algorithme de Dijkstra se révèle nécessaire car il faut réaliser un décodage de tous les chemins qui peuvent aboutir à la prédiction des mots de test et conserver celui de coût minimal. On note c_2 ce coût minimal donné par l'algorithme de Dijkstra. De même que pour c_1, le coût c_2 est supérieur au coût de la fenêtre de prédiction, par la définition de celle-ci. Par conséquent, la fonction d'élagage c_2 satisfait également la proposition 1.

Proposition 2 *La fonction d'élagage c_2 est plus efficace que la fonction c_1.*

Preuve : Le chemin emprunté lors du calcul de c_1 fait parti de l'ensemble des chemins vus lors du décodage de la fenêtre pendant l'exécution de l'algorithme de Dijkstra pour le calcul de c_2. On trouve donc $c_2 \leq c_1$ et la fonction d'élagage c_2 est plus sélective que celle définie par c_1. □

Corollaire 1 *Lorsque l'algorithme 1 donne le rang de la fenêtre de prédiction, ce rang est le bon.*

Preuve : D'après la proposition 1, tous les chemins élagués ont un coût plus élevé que le meilleur chemin qui correspond aux mots de test ; un arrêt par $is\_empty(heap)$ est donc impossible, et si on définit N_{max} de façon à ne pas atteindre la borne, l'algorithme sort avec $is\_Sol(last(Sol))$ à vrai. Cela signifie qu'alors tous les chemins d'un coût inférieur a celui des mots de test ont été prédits, et le rang obtenu est celui de la fenêtre de prédiction. □

La figure 2 montre sur un exemple l'évolution du coût de prédiction d'une fenêtre de trois mots, dans le cadre du calcul de c_1 (courbe du haut), et dans le cadre du calcul de c_2 (courbe du bas). Sur cet exemple, on note que $c_2 < c_1$. Cette différence est due à la prédiction de deux mots conjoints $[HIGH\ YIELD]$. La coupe réalisée par l'élagage est marquée des hachures ; cela signifie qu'un chemin dont le coût passe au dessus de la zone délimitée par c_2 n'est pas à intégrer dans la liste $OPEN$ de A^*. On utilisera donc la fonction $ELAGAGE = c_2$ dans l'algorithme 1 lors de la recherche R meilleurs chemins.

La borne N_{max} définissant le rang maximum pour une fenêtre de prédiction doit être déterminé empiriquement. On peut encore le placer à $N_{max} = \infty$, ainsi le corollaire 1 est toujours vérifié.

5. CONCLUSION

Dans cet article, nous avons présenté deux fonctions d'élagage pour l'algorithme A^* permettant de rechercher le rang auquel est prédit un chemin particulier d'un multigraphe. Ces fonctions se basent sur la connaissance du coût du chemin dont on cherche le rang pour éliminer les chemins de rang supérieur. L'élagage réalisé permet de limiter la combinatoire de l'algorithme. Cet algorithme A^* a été utilisé dans le cadre d'une évaluation prédictive de modèles de langage. Des distributions de rangs peuvent être comparés sur la moyenne obtenue, plus le rang moyen est faible, plus le modèle est bon prédicteur.

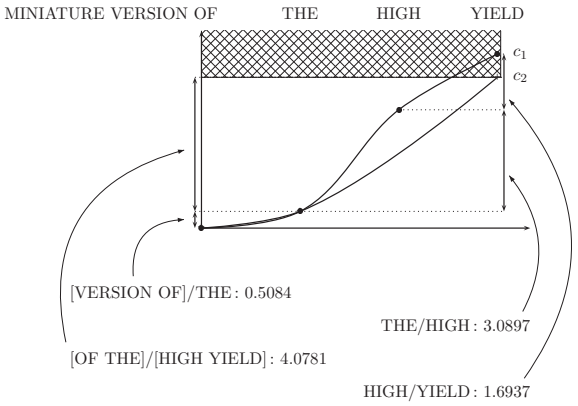


FIG. 2: Calcul des heuristiques c_1 et c_2 pour la prédiction de la fenêtre $[THE\ HIGH\ YIELD]$. c_2 est un coût plus faible que c_1, car la méthode d'estimation de c_1 ne voit pas le chemin qui prédit $[HIGH\ YIELD]$ en une seule fois, alors que l'algorithme de Dijkstra le voit lors du calcul de c_2. Un élagage peut être réalisé à partir du coût c_2 (hachures).

RÉFÉRENCES

[1] A. Bagchi and A. Manhanti. Search algorithms under different kind of heuristics : A comparative study. *Journal of the Associations for Computing Machinery*, 30 :1–21, 1983.

[2] F. Bimbot, M. El-Beze, S. Igounet, M. Jardino, K. Smaili, and I. Zitouni. An alternative scheme for perplexity estimation and its assessment for the evaluation of language models. *Computer Speech and Language*, 15(1) :1–13(13), 2001.

[3] O. Boëffard and P. Alain. Comparing rank-based statistics and standard perplexity to evaluate statistical language models. In *Proceedings of the International Conference on Speech and Computer*, pages 111–114, 2005.

[4] R. Dechter and J. Pearl. Generalized best-first search strategies and the optimality of a^*. *Journal of the Associations for Computing Machinery*, 32(3) :505–536, 1985.

[5] D. Eppstein. Finding the k shortest paths. *SIAM J. Computing*, 28(2) :652–673, 1998.

[6] V.M. Jiménez and Marzal A. Computing the k shortest paths : A new algorithm and an experimental comparison. *Lecture Notes in Computer Science*, 1668 :15–29, 1999.

[7] R. Rosenfeld. Two decades of statistical language modeling : where do we go from here ? *Proceedings of the IEEE*, 88(8) :1270–1278, 2000.

[8] C. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30 :50–64, 1951.

Etude comparative de modélisation de langage par bigrams et par multigrams pour la reconnaissance de parole

Yassine Mami[(1)], *Frédéric Bimbot*[(2)]

(1) France Télécom - 2 Avenue Pierre Marzin 22307 - Lannion - FRANCE
yassine.mami@francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
bimbot@irisa.fr

ABSTRACT

The use of stochastic ngram models has a long and successful history in the research community; nowadays ngrams are becoming quite common in oprational applications, as real-life situations demand more robust and flexible solutions. This approach is particularly interesting for its effectiveness and its robustness, but limited to modeling only local linguistic structures. To overcome this limitation, we investigate on the use of variable-length models. In this paper we consider the multigram language models and we integrate them in a speech recognition system. The experiments are carried out on a France Telecom's dialogue application for stock exchange.

1. INTRODUCTION

Le modèle de langage est parmi les plus importantes composantes dans un système de reconnaissance de parole continue. Ces modèles sont généralement des modèles de langage probabilistes et sont utilisés pour guider le décodage acoustique en apportant des contraintes linguistiques. Actuellement, les modèles de langage couramment utilisés sont les ngrams qui constituent l'état de l'art en la matière. Dans ces modèles, on estime la probabilité d'une phrase à partir des probabilités conditionnelles d'apparition d'un mot ou d'une classe de mots, étant donnés les $n-1$ mots ou classes de mots précédents. Cette approche est particulièrement intéressante pour son efficacité et sa robustesse mais limitée à la modélisation des structures linguistiques à horizon fixe. Afin de pallier les difficultés des modèles ngrams, nous proposons d'utiliser des modèles à séquences de longueurs variables [3] [5]. Les modèles multigrams font l'hypothèse qu'une phrase peut être décomposée en séquences de mots de longueur variable en faisant intervenir des probabilités sur les différentes segmentations possibles [4] [7]. Ces segmentations donnent au modèle multigram le pouvoir de capter des contraintes fortes que ne peut pas saisir un modèle ngram en raison de sa taille d'historique fixe. D'autres variantes des modèles à séquences de longueurs variables ont été proposées comme les multiclasses [10]. Cette approche est basée sur la hiérarchisation des séquences de classes syntaxiques de longueur variable. Ce type de modèle apporte aux modèles de type multigram la prise en compte de dépendances entre les séquences de classes syntaxiques de longueur variable.
Ainsi, cet article s'articule en trois parties. Dans la première, nous rappelons le principe de la modélisation par ngrams. La deuxième partie est consacrée à la modélisation par multigrams. Dans la dernière, nous présentons des évaluations comparatives de ces deux familles d'approches sur les données d'une application de dialogue dans le domaine boursier. Enfin, un ensemble de conclusions et de perspectives terminent cet article.

2. MODÉLISATION PAR NGRAMS

Le but d'un modèle de langage stochastique est d'attribuer à une séquence de mots W une probabilité $p(W)$. Soit $W = \{w_1, \cdots, w_T\}$ une séquence de mots. La probabilité de la séquence de mots s'exprime [8] :

$$p(W) = p(w_1) \prod_{i=2}^{T} p(w_i|w_1 \cdot\cdot w_{i-1}) = p(w_1) \prod_{i=2}^{T} p(w_i|h_i) \quad (1)$$

où h_i est l'historique du mot w_i.
L'approximation ngrams consiste à limiter l'historique d'un mot w_i à ses $n-1$ mots prédécesseurs, soit $h_i = w_{i-n+1} \cdots w_{i-1}$. Le modèle ngram constitue l'état de l'art actuel de la modélisation stochastique du langage. C'est un modèle stochastique pur, qui n'utilise aucune connaissance d'ordre syntaxique ou sémantique. Malgré sa simplicité de modélisation, ce type de modèle est très répandu dans les systèmes de reconnaissance vocale ; lors du décodage de Viterbi, en combinaison avec le modèle acoustique, il est assez discriminant pour favoriser certaines hypothèses de mots par rapport à d'autres et de ce fait, réduire de façon appréciable l'espace de recherche. Dans la pratique, n est rarement choisi supérieur à 3. Pour $n = 2$, la probabilité d'un mot ne dépend que du mot qui le précède ; il s'agit d'un modèle bigram :

$$P(W) \approx p(w_1) \prod_{i=2}^{T} p(w_i|w_{i-1}) \quad (2)$$

Les paramètres du modèle ngrams sont souvent estimés par maximum de vraisemblance. Pour un modèle bigram, la probabilité d'apparition du mot w_i précédé du mot w_j est :

$$p(w_i|w_j) = \frac{c(w_j w_i)}{c(w_j)} \quad (3)$$

où $c(X)$ est le nombre d'occurrences d'un événement X dans le corpus d'apprentissage.
Dans cet article, les probabilités bigrams non vues sont lissées par *discounting*. Cette technique consiste à retrancher une masse de probabilité de la masse totale accordée aux événements observés lors de l'apprentissage, masse qui se trouve ensuite redistribuée parmi les événements non observés. Plusieurs variantes de la technique de *discounting* existent : dans cet article, nous avons utilisé l'*absolute discounting* [9].

3. MODÉLISATION PAR MULTIGRAMS

Dans l'approche multigrams, une phrase est considérée comme une concaténation d'un ensemble de séquences de mots de longueurs variables [2] [5]. Soit une phrase W de T mots $W = \{w_1, \cdots, w_T\}$ et soit $S = \{s_1, \cdots, s_q\}$ une segmentation possible de la phrase W. Chaque segment est de longueur maximale m. La vraisemblance jointe de la phrase W et de la segmentation S est égale au produit des probabilités de chaque segment s_i, soit :

$$\mathcal{L}(W, S) = \prod_{i=1}^{i=q} p(s_i) \quad (4)$$

L'estimation de l'ensemble des paramètres d'un modèle multigram Θ est obtenue par maximum de vraisemblance. La vraisemblance de la phrase W est la somme des vraisemblances jointes de toutes les segmentations possibles S :

$$\mathcal{L}(W) = \sum_S \mathcal{L}(W, S) \quad (5)$$

La vraisemblance $\mathcal{L}(W)$ est une fonction non linéaire des paramètres du modèle Θ ce qui rend la maximisation directe par rapport à Θ très difficile. L'algorithme EM (*Expectation Maximisation*) permet de résoudre d'une manière élégante ce problème d'estimation relativement complexe. Au sens de cet algorithme, les données observées sont la suite des mots W et les données manquantes (ou cachées) sont les segmentations de la phrase S. L'estimation des paramètres est un processus itératif qui se déroule comme suit :

1. Initialiser les probabilités des séquences à partir des fréquences d'occurrences (cf. paragraphe 3.1).
2. Ré-estimer les probabilités des séquences multigrams (cf. paragraphe 3.2).
3. Itérer 2 jusqu'à un critère d'arrêt (la vraisemblance ne croît plus ou le nombre maximum d'itérations est atteint).

3.1. Initialisation

Soit $\mathcal{D} = \{s_1, \cdots, s_m\}$ l'ensemble des séquences multigrams formé à partir de la combinaison des mots du lexique. Les probabilités $p(s_i)$ de ces séquences peuvent être initialisées à partir de leurs fréquences d'apparition, soit :

$$p^0(s_i) = \frac{c^0(s_i)}{c^0} \quad (6)$$

où $c^0(s_i)$ est le nombre d'occurrences du segment s_i dans le corpus d'apprentissage, c^0 est le nombre total d'occurrences de touts les segments.

3.2. Estimation des probabilités des séquences multigrams

Les probabilités des séquences multigrams peuvent être apprises en utilisant l'algorithme EM ou l'algorithme Viterbi. Dans cet article, nous utilisons l'algorithme EM. La formule de ré-estimation des paramètres est alors :

$$p^{(k+1)}(s_i) = \frac{\sum_S c(s_i|S)\mathcal{L}^{(k)}(W, S)}{\sum_S c(S)\mathcal{L}^{(k)}(W, S)} \quad (7)$$

où $\mathcal{L}^{(k)}(W, S)$ est la vraisemblance de la phrase W à l'itération (k). La quantité $c(s_i|S)$ est le nombre d'occurrences de s_i dans une segmentation S, et $c(S)$ est le nombre total des séquences.
L'équation 7 est implémentée en utilisant la procédure *forward-backward*. Pour une séquence s_i de l mots, la formule de ré-estimation est :

$$p^{(k+1)}(s_i) = \frac{\sum_{t=1}^{T} \alpha_l^{(k)}(t)\beta^{(k)}(t)\delta^{s_i}_{[w(t-l+1)\cdots w(t)]}}{\beta^{(k)}(0)\gamma^{(k)}(T)} \quad (8)$$

Avec

$$\delta^{s_i}_{[w(t-l+1)\cdots w(t)]} = \begin{cases} 1 \text{ si } [w(t-l+1)\cdots w(t)] = s_i \\ 0 \text{ sinon} \end{cases}$$

La variable *forward* $\alpha(t)$ est la vraisemblance de la première partie de la phrase $W_1^t = \{w_1, \cdots, w_t\}$:

$$\alpha(t) = \sum_{l=1}^{n} \alpha(t-l)p([w(t-l+1)\cdots w(t)]) = \sum_{l=1}^{n} \alpha_l(t) \quad (9)$$

La variable *forward* $\gamma(t)$ est le nombre moyen des séquences dans une segmentation possible de $W_1^t = \{w_1, \cdots, w_t\}$:

$$\gamma(t) = 1 + \sum_{l=1}^{n} \gamma(t-l)\frac{\alpha_l(t)}{\alpha(t)} \quad (10)$$

la variable *backward* $\beta(t)$ est la vraisemblance de l'autre partie de la phrase $W_t^T = \{w_t, \cdots, w_T\}$:

$$\beta(t) = \sum_{l=1}^{n} p([w(t+1)\cdots w(t+l)])\dot{\beta}(t+l) \quad (11)$$

avec $\alpha(0) = \beta(T) = 1$, $\gamma(0) = 0$ et $1 \leq t < T$.

4. EVALUATION

4.1. Evaluation d'un modèle de langage

Si le modèle de langage est considéré comme une entité indépendante de l'application vocale, les performances sont souvent exprimées en terme de perplexité. C'est un indicateur de la capacité de prédiction du modèle de langage. La perplexité est donnée par la formule suivante :

$$PP(W) = 2^{-\frac{1}{T}\log_2 \mathcal{L}(W)} \quad (12)$$

La perplexité s'avère être un bon estimateur de la qualité d'un modèle de langage en tant qu'entité autonome. Cependant, sa valeur n'est pas forcément corrélée aux performances de ce modèle intégré dans une application de reconnaissance vocale et de son comportement en coopération avec le modèle acoustique.

4.2. Evaluation d'un système de reconnaissance de parole

Les performances d'un système de reconnaissance de parole continue sont exprimées en taux d'erreur de mots WER :

$$WER = \frac{Ins + Sub + Omi}{OK + Sub + Omi} \quad (13)$$

où Ins, Sub et Omi représentent respectivement le nombre d'insertions, le nombre de substitutions et le nombre d'omissions. OK est le nombre des mots correctement reconnus.
Nous pouvons également évaluer le système de reconnaissance en terme de précision $\mathcal{P}$ et de rappel $\mathcal{R}$ qui sont définit par [1] :

$$\mathcal{P} = \frac{OK}{OK + Sub + Ins} \text{ et } \mathcal{R} = \frac{OK}{OK + Sub + Omi}$$

Les valeurs de précision et de rappel sont combinées en une seule valeur d'évaluation en utilisant la $F\text{-}mesure$:

$$F = \frac{2 \cdot \mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}} \quad (14)$$

Le point de fonctionnement est obtenue en maximisant la $F\text{-}mesure$ ce qui permet de maximiser conjointement la précision et le rappel.

4.3. Contexte expérimental

Les expériences ont été réalisées sur une base de données interne France Télécom R&D d'une application dans le domaine boursier. Cette application de dialogue permet de gérer un portefeuille d'actions et offre les fonctionnalités suivantes :
– consultation de portefeuilles d'actions,
– consultation de carnets d'ordres,
– transactions (achats ou ventes d'actions), item et demande d'informations liées au domaine boursier.

Le domaine bancaire est un domaine très complexe, avec un vocabulaire technique spécifique et pointu. Beaucoup de recouvrements de mots existent entre les différents concepts de l'application, ainsi que des ambiguïtés pour les concepts critiques que sont les dates, les montants et les quantités. Pour cette application, le lexique est constitué de 2730 mots, dont plus de 1300 correspondent à des noms de titres d'actions sur lesquels le client peut effectuer des transactions. Le modèle de langage utilisé intègre trois classes : "ACTIONS", "MOIS" et "SOMMES". Elles représentent respectivement les noms d'actions, les mois de l'année et les sommes. Dans ce modèle, nous avons considéré que la probabilité d'appartenance d'un mois de l'année à la classe "MOIS" ou d'une somme quelconque à la classe "SOMMES" est uniforme. Par contre, trois niveaux de probabilités sont introduits au sein de la classe "ACTIONS" : le niveau de probabilité le plus fort concerne les 40 noms d'actions du CAC-40. Les deux autres permettent de classer le reste des noms d'actions.
Les données d'apprentissage se présentent sous la forme d'un corpus constitué de transcriptions de phrases prononcées par des utilisateurs de l'application de dialogue. Le corpus d'apprentissage est constitué de 24554 énoncés. Les phrases du corpus d'apprentissage sont pour la plupart des questions, des requêtes d'utilisateurs ou des commandes génériques d'annulation de requête ou de réinitialisation du dialogue, communes à toute application de dialogue. La longueur moyenne des phrases est de 6.5 mots. Le corpus de test est constitué de 1500 énoncés. La longueur moyenne des phrases est de 4.1 mots. Pour l'apprentissage des multigrams, nous avons utilisé le toolkit [6].

4.4. Evaluation de la perplexité des multigrams

Avant de tester l'intégration du modèle de langage multigram dans un système de reconnaissance de parole, nous évaluerons la perplexité du modèle multigram en fonction de m, le nombre maximal des mots dans une séquence multigram.
La figure 1 trace les variations de la perplexité pour différentes valeurs d'occurrence minimal et un nombre d'itération égal à 5. Ces deux paramètres s'avèrent peu critiques pour la perplexité. Les paramètres des multigrams ont été estimés par l'algorithme EM (équation 7). Les probabilités des séquences non vues sont lissées en apprenant ces probabilités sur le corpus constitué du corpus d'apprentissage et du lexique.

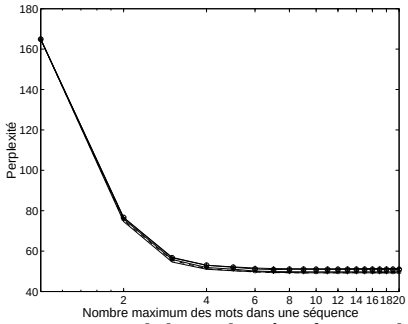


FIG. 1: Variations de la perplexité en fonction de m.

La figure 1 montre que la perplexité est maximale pour un modèle unigram (toutes les séquences ont une longueur égale à 1). La perplexité décroît très rapidement pour des valeurs du nombre maximal de mots dans une séquence inférieur à 4. Au delà de cette valeur, la perplexité varie peu et atteint une valeur minimale de 49.2 pour un nombre maximal de 5 mots dans une séquence multigrams. Au delà de 4 mots dans une séquence, le nombre d'occurrences minimal de cette séquence n'est pas un paramètre critique pour la perplexité. Par ailleurs, la perplexité du modèle bigram évaluée sur le corpus de test est égale à 32.2. Cela peut être justifié par la complexité des modèles bigrams et multigrams. En effet, le nombre de bigrams est presque trois fois supérieur au nombre d'unités du modèle multigram de taille 2. Le tableau 1 donne le nombre d'unités des autres modèles.

4.5. Performances de reconnaissance

Modèle	Nombre d'unités	WER
bigram	9719	31.0%
multigram $m = 2$	3378	37.9%
multigram $m = 3$	5280	34.8%
multigram $m = 4$	6628	34.3%
multigram $m = 5$	7523	34.0%

TAB. 1: Performances de reconnaissance ngrams vs. multigrams.

Le tableau 1 donne les taux d'erreur mots des systèmes de reconnaissance intégrant des bigrams ou des multigrams de taille des séquences $m = \{2, 3, 4, 5\}$. Le système de reconnaissance de parole intégrant les multigrams donne des meilleures performances pour des tailles de séquence plus grande, passant de $WER = 37.9\%$ pour un système multigrams de taille 2 à $WER = 34.0\%$ pour un système multigrams de taille 5. Par comparaison, le système de reconnaissance intégrant les bigrams donne les meilleures performances $WER = 31.0\%$. Cela est dû au fait que les bigrams modélisent bien la parole spontanée où les phrases ne sont pas bien structurées.

4.6. Estimation des bigrams en utilisant les séquences multigrams

Une différence substantielle entre le modèle des bigrams et le modèle des multigrams réside dans les hypothèses sous-jacentes aux dépendances existant entre les mots au sein des phrases. Le modèle bigram représente

ces dépendances sous forme d'un chaînage alors que le modèle multigram les schématise comme concaténation de blocs indépendants. Dans ce contexte, il est intéressent de considérer un modèle bigram alternatif, pour lequel l'apprentissage des probabilités est dérivé de celles des modèles multigrams. Ce modèle de bigram dérivé peut s'obtenir soit par calcul direct, soit par l'intermédiaire d'une estimation de bigrams sur des énoncés artificiels issus du modèle de multigrams (méthode de Monte-Carlo). Une première expérience dans cette direction a consisté à estimer des bigrams sur les séquences du dictionnaire de multigrams (de taille 5). Le modèle ainsi obtenu contient 5267 unités et donne un taux d'erreur de mots de 32.1%. Cette valeur, intermédiaire entre les résultats obtenus par le bigram pur et par le multigram pur, tendent à indiquer qu'une partie des erreurs causées par le multigram proviennent des contraintes structurelles du modèle. Ces résultats sont en cours de consolidation par des évaluations plus complètes et plus précises.

4.7. Diagnostic des erreurs

Pour approfondir le diagnostic entre bigrams et multigrams, nous avons comparé les types d'erreurs produits par les deux modèles sur le corpus de test. A la sortie des deux systèmes, nous obtenons 886 phrases où les sorties sont identiques et 614 phrases où les sorties sont différentes. Le taux d'erreur mots sur les 886 phrases est $WER = 16.7\%$ pour les deux systèmes. Sur les 614 autres phrases, il atteint 41.3% pour le système intégrant des bigrams et 46.3% pour celui intégrant des multigrams. Le tableau 2 présente en détail les résultats du diagnostic. Les erreurs de substitutions qui diffèrent différentes dans

bigram → multigram ↓	OK	Sub	Omi	Ins	non Ins
OK	4407	150	53	-	-
Sub	245	689	60	-	-
Omi	72	65	240	-	-
Ins	-	-	-	304	360
non Ins	-	-	-	299	-

TAB. 2: Comparaison des erreurs de reconnaissance par type d'erreur entre les systèmes à base de bigrams et de multigrams.

les deux approches correspondent souvent à des mots très différents. Quant aux erreurs d'insertions différentes sont souvent des mots courts et proches (par exemple $d' \leftrightarrow de$, $j' \leftrightarrow je$, ...). Les mots reconnus par les bigrams et omis par les multigrams sont souvent des mots courts (*de, le , la, à, pour, euh,* ...) qui servent à structurer la phrase, comme si la sortie du système multigrams était une simple suite de blocs de mots. En revanche, les mots reconnus par les multigrams et omis par les bigrams sont souvent des noms ou des verbes. La relative discordance entre les deux méthodes invite à réfléchir à l'utilisation d'une stratégie de fusion pour combiner les deux approches. En effet, si l'on calcule les performances que pourrait obtenir un système qui fusionnerait optimalement les deux approches (en ayant la connaissance du système préférable pour chaque mot décodé), on évalue (grâce à la table 2), une $F\text{-}mesure$ de 80.6%, à rapprocher des performances du bigram seul (77.4%, cf. table 3) ou du multigram seul (75.3%). Il reste donc une marge de progression de plusieurs $\%$ qui peut être en partie comblée par des méthodes hybrides ou combinées. Cela constitue un objet intéressant pour des travaux à venir.

	WER	$\mathcal{P}$	$\mathcal{R}$	$F\text{-}mesure$
bigram	31.0%	75.8%	79.0%	77.4%
multigram $m = 5$	34.0%	73.5%	77.1%	75.3%

TAB. 3: $F\text{-}mesure$ pour les bigrams vs. multigrams.

5. CONCLUSION

Dans cet article, nous avons comparé les deux approches de modélisation de langage ngrams et multigrams. Les évaluations des systèmes de reconnaissance intégrant ces deux approches montrent que l'intégration d'un modèle bigram donne des meilleures performances de reconnaissance. Cela semble être dû au fait que les bigrams modélisent bien la parole spontanée où les phrases ne sont pas bien structurées. Pour des performances similaires, les multigrams présentent l'intérêt d'avoir un nombre réduit de paramètres. L'analyse diagnostique des erreurs concordantes et discordantes pour les deux méthodes et la marge de progression qu'il met en évidence incite à explorer des approches combinées alliant la souplesse des ngrams et la capacité de structuration des multigrams.

RÉFÉRENCES

[1] F. Bimbot and G. Gravier. Evaluation des systèmes de reconnaissance de la parole. In S. Chaudiron, editor, *Traité des Sciences et Techniques de l'Information*, pages 189–213. Hermès, 2004.

[2] F. Bimbot, R. Pieraccini, and B. Atal. Variable-length sequence modeling : Multigrams. In *IEEE Signal Processing Letters*, volume 2, pages 111–113, 1995.

[3] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. Modèles de séquences à horizon variable : Multigrams. In *JEP*, 1994.

[4] S. Deligne. *Modèles de séquences de longueurs variables, application au traitement du langage naturel et de la parole*. PhD thesis, ENST, 1996.

[5] S. Deligne and F. Bimbot. Language modeling by variable length sequences : Theoretical formulation and evaluation of multigrams. In *ICASSP*, pages 169–172, 1995.

[6] S. Deligne and F. Bimbot. *Multigram Package*. ENST, 1997.

[7] S. Deligne and F. Bimbot. Learning a syntagmatic and paradigmatic structure from language data with a bi-multigram model. In *International Conference on Computational Linguistics*, pages 300–306, 1998.

[8] M. Federico and R. De Mori. Spoken dialogue with computers. *Academic Press*, pages 202–210, 1998.

[9] I. H. Witten and T. C. Bell. The zero-frequency problem : estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 1991.

[10] I. Zitouni, K. Smaili, J.-P. Haton, S. Deligne, and F. Bimbot. A comparative study between polyclass and multiclass models. In *ICSLP*, 1998.

La prosodie des mots grammaticaux : le cas des deux déterminants "du" et "deux" dans l'apprentissage du français langue étrangère

KAMIYAMA Takeki

ILPGA, Université de la Sorbonne Nouvelle - Paris III
19, rue des Bernardins 75 005 Paris, France
takekik@phiz.c.u-tokyo.ac.jp
http://www.cavi.univ-paris3.fr/ilpga/ed/student/stkt/

ABSTRACT

Does explicit knowledge of prosody help L2 learners to identify the two determiners "du" and "deux" in French? An analysis of 162 sentences read by 3 French native speakers show the expected tendency of F0 and duration ("deux" being longer and higher than the function word "du"). Then, 3 sets of 8 synthesised stimuli were generated using Mbrola, with expected and unexpected f0 and duration patterns. A perception experiment with 16 French native speakers suggests that they tend to be biased by the unexpected prosody (duration, in particular) when they listen to the sentences with white noise. In another experiment, three groups of Japanese-speaking learners were asked to identify the two words in 48 sentences read by a native speaker. The preliminary results suggest that teaching explicit knowledge of prosody might facilitate the acquisition.

1. INTRODUCTION

Le changement phonétique montre une interaction importante entre le segmental, l'accent lexical, le statut grammatical de mots, et la prosodie (rythme et intonation) dans les langues du monde. La place de la prosodie dans l'enseignement de la prononciation des langues étrangères a été relativement marginale, malgré l'existence de contributions importantes telles que le manuel emblématique de O'Connor et Arnold (1961/1973) sur l'intonation de l'anglais britannique. Dalton & Seidlhofer (1994: 73) mentionne la "teachability" relativement élevé du segmental et l'importance communicative relativement élevée de l'intonation dans l'enseignement de la prononciation de l'anglais. Selon elles, l'accent lexical (stress) est situé dans une zone de "maximum overlap" des deux facteurs. Quant à l'enseignement du français, Wioland (1991) propose la dernière syllabe de groupes rythmique en tant que la position favorable pour enseigner et apprendre des contrastes au niveau segmental. Il a été montré que la prosodie joue un rôle majeur quand on entend la parole dans une condition défavorable (dans un bruit de fond, par exemple), et dans la parole synthétisée.

Dans la présente étude, nous prenons le cas des deux déterminants, le partitif *du* et le numéral *deux*, afin d'illustrer l'interaction entre le segmental et la prosodie dans l'enseignement et l'apprentissage du français langue étrangère. Les deux mots en question sont des déterminants, et donc peuvent occuper la même position syntaxique, c'est-à-dire (immédiatement ou non) avant un substantif. Nous avons ainsi des paires minimales de phrases telles que « *nous avons du chocolat* » et « *nous avons deux chocolats* ». Sur le plan segmental, la seule différence qui se trouvent est la différence de voyelle : le /y/ est une voyelle antérieure fermée arrondie, avec le F2 et le F3 proches (CALLIOPE 1989: 84), tandis que le /ø/ est mi-fermée, avec les formants plus ou moins équidistants (figure 1). Il est connu que l'identification et la discrimination des voyelles antérieures arrondies sont difficiles pour les locuteurs des langues qui ne les possèdent pas dans leur système phonémique (Gottfried 1984, Levy & Strange 2002). C'est le cas du japonais, qui est un système à 5 voyelles.

Cependant, le timbre vocalique n'est pas la seule différence qui permet de distinguer les réalisations de ces deux mots. En français, comme dans beaucoup d'autres langues, les mots grammaticaux ont tendance à se prononcer avec un f0 plus bas et une durée plus courte, comme l'observe Vaissière (1980). Dans nos exemples, le partitif *du* est un mot grammatical, tandis que le numéral *deux* est considéré comme un mot lexical qui peut être en contraste avec d'autres numéraux.

Si ces différences prosodiques sont acoustiquement et perceptivement pertinentes dans ces phrases, et que les apprenants en sont conscients, pourraient-ils distinguer les deux déterminants plus facilement ? Afin d'examiner cette question, les expériences suivantes ont été effectuées : 1) analyse acoustique de phrases qui contiennent les deux déterminants *du* et *deux*, lues par des locuteurs natifs du français, 2) expérience perceptive auprès des auditeurs français, 3) expérience perceptive auprès des apprenants japonophones.

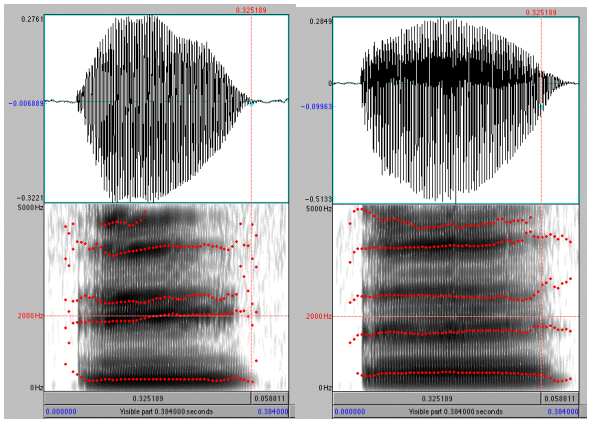


Figure 1: L'oscillogramme et le spectrogramme à bande large (fenêtre : 5 ms) des voyelles /y/ (à gauche) et /ø/ (à droite) prononcées en isolation par la locutrice 3.

2. EXPERIENCE 1: LA PRODUCTION DE LA CONTRASTE "*DEUX*"-"*DU*" PAR LES LOCUTEURS NATIFS

2.1. Procédures

3 locutrices natives du français (étudiantes doctorantes) qui vivent dans la région parisienne ont lu 162 phrases qui contiennent un syntagme nominal avec *du* ou *deux*, soit en tant qu'attribut ("*C'est du/deux thé(s).*"), soit en tant que complément d'objet direct. La longueur du syntagme nominal sujet varie entre 1 ("*nous*", "*Jean*") et 6 ("*le garçon du village*"), celle du syntagme verbal entre 2 ("*avons*") et 4 ("NP *a commandé*"), celle du syntagme nominal objet entre 2 ("*du/deux thé(s)*") et 4 ("*du/deux chocolat(s)*"). Les phrases ont été présentées aux locuteurs une par une sur un écran d'ordinateur. Elles ont été mises dans un ordre semi-aléatoire prédéfini, tout en évitant des séquences où l'un des déterminants apparaisse trop fréquemment. La durée et la fréquence fondamentale moyenne des segments ont été mesurées sur Praat.

2.2. Résultas

La voyelle dans le mot *deux* s'avère significativement plus longue (91 ms, 76 ms, 88 ms pour chaque locuteur) que celle de *du* (72 ms, 72 ms, 66 ms : figure 2). Cette différence est statistiquement significative (test-t non apparié de Student : t160 = 7,36, 2,06, et 7,98 respectivement pour chaque locutrice. $p < 0,05$ pour toutes les trois locutrices). Également, le f0 est plus élevé dans la voyelle de *deux* (222 Hz, 200 Hz, 229 Hz) que dans celle de *du* (192 Hz, 178 Hz, 178 Hz: figure 3). Cette différence est également statistiquement significative (test-t non apparié de Student : t139 = 17,74, t156 = 14,64, t160 = 18,77 respectivement pour chaque locutrice. $p < 0.05$ pour toutes les trois locutrices. Le nombre inégal de degré de liberté est dû aux cas où le F0 n'a pas été détecté.). Voir la figure 1 pour les tendances générales des phrases.

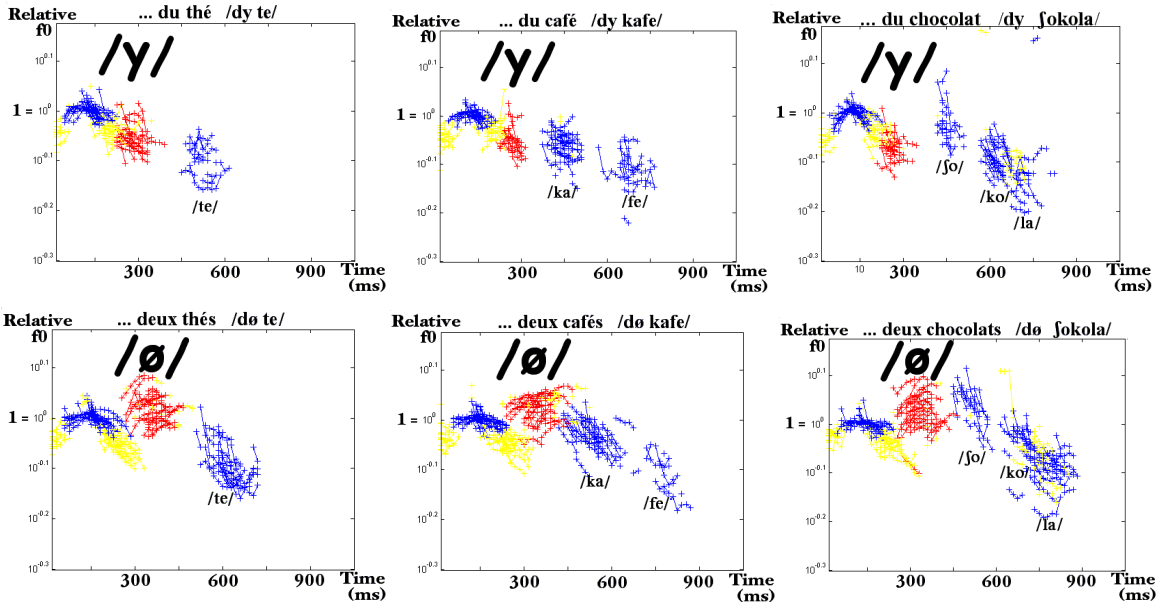


Figure 1 : Les courbes de f0 superposées (valeurs relatives à la moyenne de la voyelle qui précède /dy/ et /dø/, sur une échelle logarithmique) des 27 phrases qui se terminent respectivement par "*du thé / café / chocolat*" (en haut) et "*deux thés / cafés / chocolats*" (en bas) lues par la locutrice 1. Les croix rouges représentent les voyelles dans les deux mots en question.

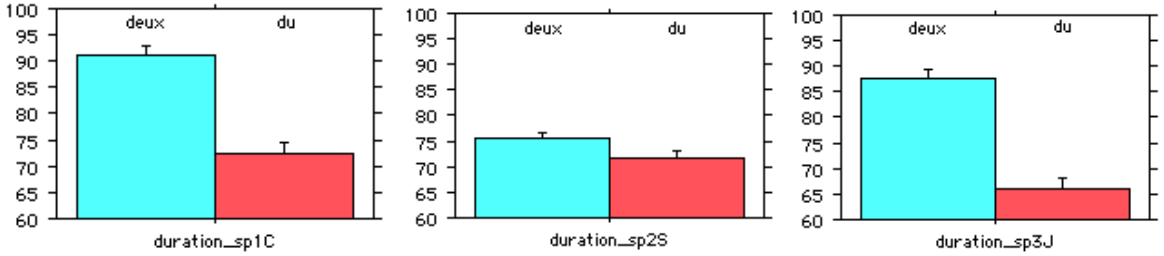


Figure 2 : La durée moyenne (ms) de la voyelle des deux déterminants *deux* et *du* prononcés par 3 locutrices natives (dans 81 phrases pour chacun des deux déterminants). La barre d'erreur représente 1 écart type.

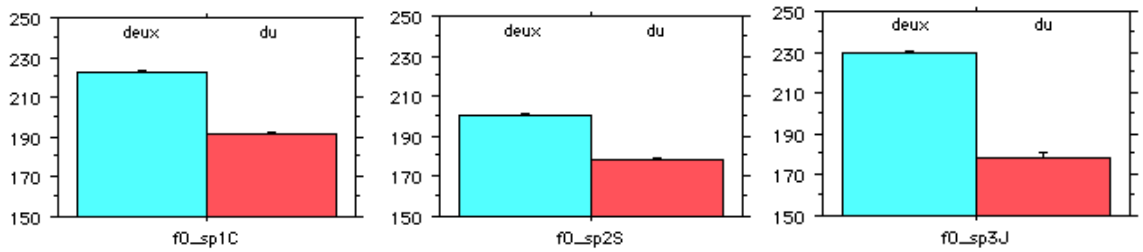


Figure 3: Le f0 moyen (Hz) durant la voyelle des deux déterminants *deux* et *du* prononcés par 3 locutrices natives (dans 81 phrases pour chacun des deux déterminants). La barre d'erreur représente 1 écart type.

3. EXPERIENCE 2 : LA PERCEPTION DES LOCUTEURS NATIFS

Afin d'étudier l'aspect perceptif des tendances observées dans la partie précédente, une expérience de perception a été effectuée auprès des locuteurs natifs. Les locuteurs natifs se servent-ils des informations prosodiques concernant ces deux déterminants quand les segments ne sont pas facilement perceptibles ?

3.1. Stimuli

Les stimuli ont été générés par le synthétiseur Mbrola, en utilisant une base de donnée de segments d'une locutrice francophone (*fr 4*). Pour chaque paire de phrases (ex. "*Le garçon a commandé du/deux chocolat(s)*"), les stimuli diffèrent uniquement par la voyelle (/y/ et /ø/), le f0 et la durée des deux déterminants. Le pattern de f0 et de durée des phrases imitent la production de la locutrice 1 (Figure 1). En se fondant sur ces phrases, 4 conditions ont été crées : 1) aucune modification, 2) durée inattendue (ex. *du* avec une durée attendue de *deux*), 3) f0 inattendu, 4) durée et f0 inattendus. 3 paires de phrases, chacune se terminant par "*chocolat*", "*café*", "*thé*", ont été choisies, ce qui fait 24 stimuli (2 déterminants * 3 paires * 4 conditions). Un bruit blanc qui a approximativement la même amplitude que le sommet d'amplitude des stimuli a été ajouté aux stimuli afin de créer la condition bruitée (48 stimuli).

3.2. Procédures d'exprérience

16 locuteurs natifs du français habitant dans la région parisienne ont participé à l'expérience, composée de deux parties. Dans la première, les stimuli avec le bruit blanc ont été présentés, tandis que dans la seconde, c'est les stimuli sans bruit. L'ordre de présentation des deux parties a été changé pour la moitié des auditeurs. Les stimuli, précédés par un bip sonore (440 Hz, 50 ms), ont été présentés dans un ordre semi-aléatoire pré-établi, et la liste des stimuli ont été répétée deux fois (trois fois pour la condition bruitée) dans des ordres différents. La tâche des auditeurs consistait à écouter le stimulus et à répondre à la question suivante en cliquant sur la case correspondant à leur perception : "*Avez-vous entendu '... deux N' ou '... du N' ?*"

3.3. Résultats

Les auditeurs français ont identifié les deux mots "correctement" presque parfaitement (sauf 1 sur 48 occasions de réponses * 16 auditeurs) sans bruit, quel que soit le pattern prosodique, attendu ou non. Quant aux stimuli bruités, aucun des facteurs suivants a contribué à un effet statistiquement significatif sur les résultats : la position des deux boutons de réponse, l'ordre de présentation (avec/sans bruit), l'ordre des stimuli sur la liste. La durée serait le facteur le plus important qui détermine le jugement des auditeurs : la phrase "*Le garçon a commandé du chocolat*" avec la durée attendue de *deux* a été jugée comme contenant *deux* dans 22 sur 48 occasions de réponses, et "*Le garçon a commandé deux cafés*" avec la durée de *du* a été considérée comme comportant *deux* dans 20 sur 48 occasions. En revanche, en combinant avec le facteur f0, cette tendance n'a pas toujours été observée clairement. Ce résultat suggère l'importance de la durée, un facteur largement négligé dans les études récentes (par rapport au f0).

4. EXPERIENCE 3 : L'IDENTIFICATION PAR LES APPRENANTS JAPONOPHONES

Après avoir observé la perception des locuteurs natifs, qui seraient influencés par la prosodie attendue dans la condition bruitée, une autre expérience a été effectuée afin d'étudier la perception des apprenants japonophones. Dans ce test, seuls les stimuli naturels, lus par une locutrice native, ont été utilisés.

4.1. Sujets

14 locuteurs natifs du japonais apprenant le français langue étrangère/seconde, qui suivaient les cours de français oral à l'ILPGA (Institut de phonétique et linguistique générales et appliquées), Université de Paris III, ont participé au test. Leur expérience d'apprentissage variait de 9 mois à 4 ans.

4.2. Procédures d'expérience

Les apprenants ont écouté 48 phrases lues par la locutrice 1 dans l'expérience 1. La présente expérience consistait de 3 parties, précédées par une session d'entraînement : 1) pré-test, 2) traitement, 3) post-test.

Les sujets y ont participé en groupes de 2 à 4, et ils ont reçu 3 types de traitement différents selon le groupe. 1) Aucun retour explicite : ils ont eu la réponse des 6 phrases utilisées dans la session d'entraînement. 2) Description explicite des différences articulatoires et perceptives sur les voyelles /y/ et /ø/, ainsi que la réponse de la session d'entraînement. 3) Description explicite sur les patterns prosodiques fréquemment observés (*deux* plus long et plus haut), ainsi que la réponse de l'entraînement. Les pré-test et le post-test étaient exactement identiques. Dans chacun des deux, la liste des stimuli a été répétée deux fois dans des ordres différents, avec une pause entre les deux. La tâche des apprenants consistait à écouter le stimulus et à répondre à la question suivante en entourant leur choix sur une feuille de réponse : "*Avez-vous entendu « du » ou « deux » ?*" Nous avons considéré que les apprenants ont répondu correctement si et seulement s'ils étaient consistants dans leur réponse, c'est-à-dire, s'ils ont donné la bonne réponse deux fois au même stimulus, avant et après la pause.

4.3. Résultats préliminaires

Quel que soit le type de traitement, la plupart des apprenants a donné plus de 40 réponses correctes (sur 48 phrases) dans le pré-test. Même s'il y a une légère amélioration dans le post-test (2-4 réponses correctes de plus), nous devrions considérer l'effet de plafond. Il y a cependant 2 apprenants qui ont donné respectivement 30 et 32 réponses correctes dans le pré-test, et amélioré leur score jusqu'à 40 et 43 respectivement dans le post-test, après avoir reçu des informations explicites sur les différences prosodiques.

5. CONCLUSIONS

Les résultats de l'expérience 1 montrent que les locuteurs natifs du français ont tendance à prononcer le numéral *deux* avec une durée plus longue et un f0 plus élevé que le partitif *du*. Ceux de l'expérience 2 suggèrent que les locuteurs natifs seraient influencés, lors de l'identification des deux déterminants en question, par le pattern prosodique attendu (au moins par la durée) dans un environnement défavorable, même si cette influence n'est pas observée quand les informations segmentales sont accessibles facilement (sans bruit).

Afin de valider les tendances observées, il faudra effectuer davantage d'expériences, en utilisant des stimuli et des conditions améliorés. L'expérience 3 pourrait mener ultérieurement aux études qui permettraient de confirmer que les connaissances explicites en prosodie facilitent la compréhension des mots qui comportent des segments difficiles, mais les données préliminaires de la présente étude ne permettent pas de valider suffisamment cette hypothèse. Cette expérience devra être effectuée auprès des apprenants moins avancés qui ont eu moins d'input sonore de la langue cible.

Notre expérience suggère également que ce serait avantageux d'apprendre aux apprenants à identifier des contrastes segmentales dans les conditions défavorables, pour qu'ils puissent être plus sensibles aux indices secondaires qui sont souvent essentiels dans la vie quotidienne. L'utilisation de la parole synthétisée, qui permet de contrôler chaque paramètre (synthétiseurs à formants nécessite le contrôle des distances entre les formants, tandis qu'avec Mbrola nous ne pouvons contrôler que la durée et le f0), pourrait s'avérer essentielle pour que les apprenants soient plus sensibles aux indices acoustiques qui ne sont pas utilisés dans leur langue native.

Même si le nombre des paires comme *du* et *deux* est très limité, ceci pourrait être un exemple qui montre l'interface entre le segmental et la prosodie dans l'enseignement et l'apprentissage de la prononciation des langues étrangères.

BIBLIOGRAPHIE

[1] P. Boersma, D. Weenink. *Praat: doing phonetics by computer* (Version 4.3.00) [logiciel informatique]. Tiré le 26 février 2005, de http://www.praat.org/, 2005.

[2] CALLIOPE. *La parole et son traitement automatique*. Masson, Paris, Milano, Barcelona, Mexico, 1989.

[3] C. Dalton, B. Seidlhofer. *Pronunciation*. Oxford University Press, Oxford, 1994.

[4] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. van der Vrecken. The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96, Philadelphia*, pages 1393-1396 (http://tcts.fpms.ac.be/synthesis/mbrola.html), 1996.

[5] T. L. Gottfried. Effects of consonant context on the perception of French vowels. *Journal of Phonetics*, 12: 91-114, 1984.

[6] E. S. Levy, W. Strange. Effects of consonantal context on perception of French rounded vowels by American English adults with and without French language experience. *Journal of the Acoustical Society of America*, 111(5): 2361-2362, 2002.

[7] J. D. O'Connor, G. F. Arnold. *Intonation of Colloquial English*. Longman, London, 1961/1973.

[8] J. Vaissière. La structuration acoustique de la phrase française. *Annali della Scuola Normale superiore di Pisa*, Classe di lettere e filosofia, Serie III, X,2 : 529-560, 1980.

[9] F. Wioland. *Prononcer les mots du français*. Hachette, Paris, 1991.

Aspects phonologique et dynamique de la distinctivité au sein des systèmes vocaliques: une étude inter-langue

Christine Meunier, Robert Espesser, Cheryl Frenck-Mestre

Laboratoire Parole et Langage – C.N.R.S. U.M.R. 6057
Université de Provence, Aix-en-Provence, France
christine.meunier@lpl.univ-aix.fr
http://www.lpl.univ-aix.fr/~meunier/

ABSTRACT

This study is a cross-linguistic investigation of qualitative and quantitative variations due to 1/ the structure of vocalic system, 2/ the amount of context within speech message. We hypothesize that phonetic distinctivity of vowels in a language is relative to 1/ the properties of the phonological system, 2/ the amount of informational context. Three languages (Spanish, French and English) were analyzed in three different types of speech (isolated vowels, within words and within texts). Results show 1/ centralization in the three vocalic systems relative to the amount of context, 2/ an increase of vowel dispersion also due to an increase of context information.

1. INTRODUCTION

Les systèmes vocaliques des langues du monde ainsi que la logique de leur structure ont été finement décrits (Lindblom [1], Schwartz et al. [2]). La structure de l'inventaire des voyelles dans une langue n'est pas aléatoire mais suit une logique universelle relative aux contraintes articulatoires et à la distinctivé nécessaire au sein de chaque système. Ainsi, si un système ne présente que trois voyelles, ces voyelles occuperont les espaces articulatoires les plus éloignés.

Si l'inventaire et la structuration du système (donc sa forme) suivent cette logique, on peut faire l'hypothèse que les réalisations des voyelles (donc la 'matière') vont dans le sens de cette distinctivité. Il a ainsi été montré que les locuteurs d'une langue tendent à utiliser un 'hyper-espace' de façon à rendre les voyelles très distinctes les unes des autres (Johnson [3]). Certains travaux ont montré que la densité d'un système (le nombre plus ou moins grand de voyelles) pourrait avoir une influence sur la dispersion des réalisations: plus le nombre de voyelles est réduit, plus la dispersion serait importante (Manuel & Krakow [4]). Là encore la nécessité du caractère distinct du système renforcerait cette hypothèse. Toutefois, plusieurs études ont produit des résultats inverses: dans un système à 3 voyelles (Ami), Maddieson [5] a pu montré que les réalisations n'était pas précisément plus dispersées. De même, dans une étude récente (Meunier et al. [6]), nous avons observé une plus grande dispersion en Anglais (12 voyelles) qu'en Espagnol (5 voyelles). Nous avions alors conclu qu'il n'est pas exclu que la densité joue un rôle dans la dispersion des réalisations, mais qu'elle représente probablement un des nombreux facteurs qui conditionnent la réalisation des voyelles.

Dans certaines situations, une tendance à la réduction du système vocalique lui-même est observée. Lindblom [7] observent que les voyelles inaccentuées du suédois se centralisent lorsque le débit augmente. La persistance du phénomène de centralisation dans les systèmes vocaliques français et allemand (et non seulement spécifique aux systèmes à accent fort) a conduit Gendrot & Adda-Decker [8] à conclure qu'il s'agirait autant d'un phénomène purement physiologique (réduction de l'effort articulatoire) que d'un phénomène du à des contraintes linguistiques.

Notre hypothèse est que l'argument physiologique ne peut être une explication exhaustive du phénomène. La centralisation du système entraîne de fait une réduction de la distinctivité globale (moins de distance entre chaque élément). Cette réduction n'est possible que si d'autres aspects de la communication apporte une compensation et donc maintiennent la distinctivité à un autre niveau. Cette *variabilité* est donc *adaptative* (*H & H Theory*, Lindblom [9]). Mais plus qu'un constat d'une hypo-articulation en parole non contrôlée, il nous semble nécessaire de rendre explicite la dynamique des mouvements de variations. Nos productions sonores participent d'un équilibre entre efficacité articulatoire et préservation de la communication. Comment fonctionne cet équilibre dans le système de la langue? Notre hypothèse est qu'il est fonction d'au moins deux dimensions: l'une, que nous nommerons *Propriétés Statiques* (PS), est régie par la structure du système de sons lui-même (inventaire et propriétés du système phonologique), l'autre, nommée *Agencement Dynamique* (AD), est régie par l'ajustement ponctuel du degré d'information dans les différents secteurs de la communication (l'information présente dans d'autres secteurs linguistiques autorise une hypo-articulation). Notre objectif est de rendre explicite l'interaction entre ces deux dimensions.

2. MÉTHODE

Notre objectif étant d'évaluer les mouvements de variation des voyelles selon, d'une part, la propriété des systèmes et, d'autre part, le degré d'information contextuel, notre matériel est donc multilingue et basé sur différents types de corpus.

Trois langues sont étudiées: le français (FR), l'anglais (EN) et l'espagnol (SP). L'anglais distingue entre 13 et 15 voyelles orales alors que le français en distingue entre 10 et 12. En ce sens, la densité des deux systèmes est assez semblable, mais l'organisation du système des deux langues est assez différente. Les voyelles orales du système français se distinguent aisément avec les seuls indices F1 et F2. En

revanche, le système de l'anglais possède des indices secondaires (Vallée [10], Schwartz et al. [2]) tels que la durée et l'accent qui rendent les indices F1/F2 insuffisants pour l'identification des voyelles (Meunier et al [6]). L'espagnol comporte un inventaire comparativement moins fourni, avec seulement 5 voyelles, mais dans cet inventaire, nous trouvons les mêmes voyelles qu'en français et en anglais: /a/, /e/, /o/, /i/, /u/. Ces trois langues offrent la possibilité de distinguer l'effet de la densité (inventaire plus ou moins fourni), de l'effet de la complexité du système (indices primaires/secondaires).

Trois locuteurs (2 femmes et 1 homme) de chaque langue ont été enregistrés dans trois types de corpus différents.

- Voyelles prononcées isolément (ISO) : ce corpus permet d'évaluer la variabilité des voyelles hors contexte. Il s'agit d'une situation de production très contrôlée qui, normalement, occasionne peu de variation.
- Voyelles insérées dans des mots monosyllabiques prononcées isolément (WORD) : ce corpus permet d'évaluer la variabilité vocalique au sein du lexique. Il permet en outre d'observer des effets de coarticulation. Cette situation de production est plus naturelle et devrait occasionner une variabilité plus importante.
- Les mots du corpus WORD sont insérés dans deux textes (TXT) : ce corpus permet d'évaluer la variabilité au sein d'un contexte varié et non contrôlé. Le débit est bien sûr plus important dans ce type de corpus.

Si l'ensemble des analyses représente 3600 voyelles mesurées, le nombre de voyelles caractéristiques de chaque condition (langue, corpus, locuteur, etc.) est assez faible et limite la portée de nos résultats qui, en conséquence, nous informe essentiellement sur des tendances. Le nombre d'observations de chaque voyelle et le suivant:
- corpus ISO (5 répétitions): 10 valeurs par voyelle/langue/locuteur (il y a donc 10 mesures de /i/ pour le locuteur français LJ dans le corpus ISO)
- corpus WORD (5 rép.): 30 valeurs par voy./langue/loc.
- corpus TXT (4 rép.): 8 valeurs par voy./langue/loc.

Les analyses ont porté sur la mesure de F1 et F2 (détection automatique des formants à l'aide du logiciel ESPS Entropics puis vérification manuelle). Deux types de variations sont observées: les variations qualitatives et les variations quantitatives (fig. 1).

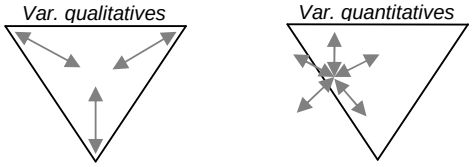


Figure 1: Type de variations observables.

Nous appelons *variations qualitatives* les variations marquées par un déplacement des cibles phonétiques. Ces variations seront observées concernant les effets de centralisation des systèmes: les valeurs moyennes se déplacent vers le centre du système. Les *variations quantitatives* caractérisent la dispersion des voyelles autour des valeurs moyennes de chaque voyelle.

L'intérêt pour nous de distinguer ces deux types de variations est d'obtenir une évaluation de la distinctivité des systèmes en fonction de différents facteurs. Par exemple, un système dont les réalisations sont centralisées et marquées par une forte dispersion des valeurs sera considéré comme peu distinctif. Cette distinctivité est évaluée en fonction de l'inventaire des systèmes (PS) et du degré de contextualisation (AD).

3. RÉSULTATS

3.1. Variations qualitatives : centralisation des systèmes

Concernant le phénomène de centralisation, nous avons porté notre attention sur les voyelles présentes dans les trois langues /i/ /e/ /a/ /o/ /u/ considérant qu'elles étaient suffisantes pour représenter fidèlement le mouvement vers le centre du système. Notons d'emblée que nos données sont hétérogènes, cela pour deux raisons principales: 1/ une importante variabilité inter-locuteur, 2/ un faible nombre d'observations pour chaque voyelle dans chaque condition. De ce fait, nous n'avons marqué significatifs (* inscrits sur les fig. 2 à 4) que les effets de variations significatifs pour chacun des trois locuteurs.

Variations qualitatives de F1 (fig. 2): on note essentiellement la fermeture des systèmes espagnols et français (F1 de /a/ diminue progressivement). Cette différence est significative pour les 3 locuteurs de chacune des deux langues (p<, 0010). Pour l'anglais, les variations ne vont pas dans le même sens pour tous les locuteurs.

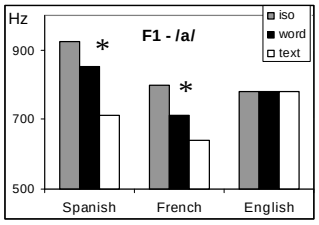


Figure 2: valeurs moyennes du F1 (en Hz) de /a/ dans les trois langues et dans les trois corpus

L'augmentation de F1 pour les voyelles fermées est une tendance qui n'est pas systématique (significative seulement pour certains locuteurs). Il est probable que cette tendance pour les voyelles fermées se confirme avec un nombre de données élargi. Notons toutefois qu'avec peu de données, la fermeture des systèmes FR et SP est bien nette.

Variations qualitatives de F2 (fig. 3-4): la centralisation de F2 est systématique pour le système espagnol aussi bien pour les voyelles antérieures (/i/ /e/) que pour les voyelles postérieures (/u/ /o/). En français, elle n'est systématique pour tous les locuteurs que pour les voyelles mi-ouvertes (diminution de F2 pour /e/, et augmentation de F2 pour /o/). En revanche, pour les voyelles fermées, la tendance est moins claire: très peu marquée pour /i/ et bien marquée pour /u/ mais seulement pour le corpus TXT. En anglais, seule l'augmentation de F2 de /o/ est significative pour tous les locuteurs. Là encore, il est probable que la tendance à la centralisation devienne plus nette avec un plus grand nombre

de données. De nouveau, le système espagnol semble plus sensible à la modalité de parole pour la centralisation.

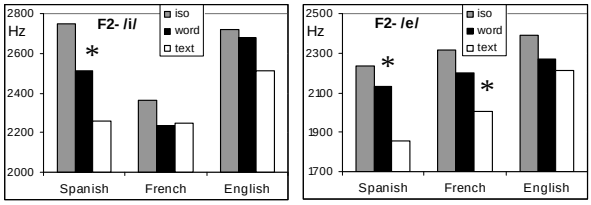


Figure 3: valeurs moyennes du F2 (en Hz) des voyelles antérieures (/i/ /e/) dans les trois langues et les trois corpus.

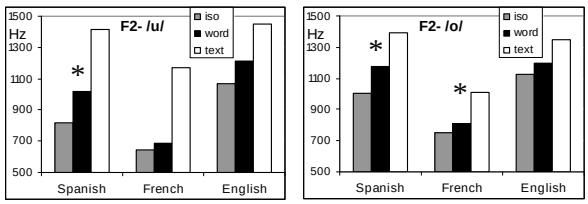


Figure 4: valeurs moyennes du F2 (en Hz) des voyelles postérieures (/u/ /o/) dans les trois langues et les trois corpus.

En résumé, on observe une tendance générale vers la centralisation des systèmes avec des effets plus marqués en fonction 1/ du système phonologique (SP>FR>EN); 2/ du type de corpus (TXT>WORD>ISO). Dans les trois langues, le système vocalique se réduit proportionnellement à la contextualisation, mais cette réduction est relative à l'inventaire phonologique de chaque langue (fig. 7 et 8).

3.2. Variations quantitatives : dispersion des voyelles

Pour obtenir une idée globale de la dispersion moyenne des voyelles dans un système, nous avons recueilli les Déviations Standards (SD) des valeurs de F1 et F2 de chaque voyelle, pour chaque locuteur et dans chaque condition. Nous avons ensuite calculé la moyenne de ces SD pour chaque langue. Ainsi, une valeur élevée sur ces figures représente une forte dispersion des voyelles.

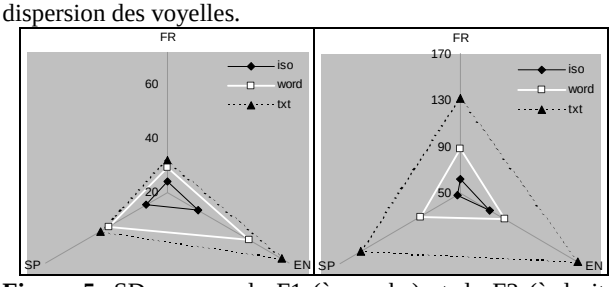


Figure 5: SD moyens du F1 (à gauche) et du F2 (à droite) des voyelles des trois langues (SP, FR et EN).

L'effet principal observé est une augmentation des valeurs des SD en fonction du degré de contextualisation (fig. 5). Plus le contexte est important (moins la situation de production est contrôlée), plus les réalisations phonétiques sont dispersées. On constate un effet grandissant de la dispersion proportionnel à la quantité de contexte: ISO>WORD>TXT. On constatera que pour les productions anglaises, le degré de dispersion est systématiquement plus grand que pour le français et l'espagnol, même si l'effet 'corpus' reste le même dans les trois langues.

3.3. Distinctivité des systèmes

Nous avons également calculé la moyenne de toutes les mesures de voyelles effectuées sur un système. Les SD de ces valeurs nous donnent une idée de la dispersion ou de la centralisation de l'ensemble du système (fig. 6). Nous observons ici que la dispersion de chaque système est plus grande en corpus ISO et plus faible en corpus TXT. Ce qui confirme nos observations sur les variations qualitatives. On remarque que, si la tendance reste la même, les systèmes espagnols et anglais évoluent différemment: les productions anglaises montrent un espace élargi en TXT qui s'accentue peu en ISO; tandis que les productions espagnoles sont très centralisées en TXT et extrêmement périphériques en ISO.

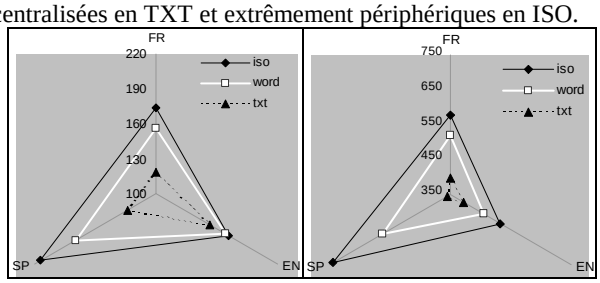


Figure 6: SD moyens du F1 (gauche) et du F2 (à droite) de l'ensemble des valeurs de chaque système.

En comparant les figures 5 et 6, on constate un effet inverse des dispersions: en TXT, dispersion maximale de chaque voyelle (fig.5) dans un espace réduit au maximum (fig. 6). Ces résultats confirment une *hyper-distinctivité* hors contexte (ISO) et une *hypo-distinctivité* en contexte (TXT). L'importance de la distinctivité est pondérée par la nature de chaque système.

4. DISCUSSION

Les résultats obtenus dans cette étude tendent à montrer que la distinctivité phonétique effective au sein des systèmes vocaliques est fonction 1/ des propriétés du système (son inventaire et le système d'indices qui le caractérise), 2/ de la quantité de contexte spécifique à la situation de production.

4.1. L'effet 'Langue'

Dans des travaux antérieurs, nous avions constaté que la densité des systèmes ne semblait pas jouer de rôle en production (Meunier et al. [6]). Ces travaux portaient sur l'observation du corpus ISO et les locuteurs espagnols montraient un hyper-espace en production et de très faibles dispersions des réalisations. Or si l'on regarde maintenant les productions en contexte, il semble que la densité joue un rôle important: le système espagnol se réduit considérablement par rapport à celui de l'anglais (fig. 7 et 8). Il semble donc que *moins un système est dense, plus il est élastique*. D'une certaine façon, le système de l'anglais est beaucoup moins sensible à la situation de production (résistance à la centralisation, dispersion déjà très importante en ISO, etc). Quand au système français il semble intermédiaire, ce qui pourrait être du à la nature de son système: aussi dense que le système anglais mais moins complexe concernant le type d'indices nécessaires à la distinction des voyelles.

4.2. L'effet 'corpus'

Au travers de cette étude, les effets des corpus sont assez nets et confirment une hyper-articulation des voyelles en corpus contrôlé (ISO) et une hypo-articulation en parole contextualisée (TXT).

L'effet systématiquement intermédiaire du corpus WORD n'est pas sans intérêt. En effet, les travaux de Lindblom [7] et de Gendrot & Adda-decker [8] tendaient à montrer que la centralisation était corrélée avec l'augmentation du débit. Nos premiers résultats concernant la durée (non exposés ici) montrent qu'il n'y a pas de différence entre les durées des voyelles ISO et celles du corpus WORD (en français, les durées WORD sont même plus longues). En revanche, la position intermédiaire de ce corpus pourrait trouver son origine dans la quantité d'information intermédiaire de l'unité lexicale.

4.3. La régulation de la distinctivité

D'une certaine façon, l'observation d'une hypo-articulation en parole non contrôlée, est juste la confirmation d'une évidence. Il va de soi que plus la parole est contextualisée, plus le débit augmente, et plus les cibles articulatoires ne peuvent être atteintes. En revanche, il nous semble intéressant de s'interroger sur des causes et le fonctionnement de cette évidente hypo-articulation. La première cause qui nous vient à l'esprit est mécanique : la réduction vocalique serait liée à l'optimisation des contraintes gestuelles. Toutefois, cette contrainte ne peut expliquer à elle seule la diminution de la distinctivité au sein des systèmes. Cette diminution n'est possible que si l'intelligibilité du message reste intact, et donc si l'apport d'information est assuré à un autre niveau. Il semble impossible de distinguer les contraintes mécaniques des contraintes informationnelles ; elles vont de pair et l'une conditionne l'autre. Comment expliquer l'hyper-articulation en production de voyelles isolées sinon par une recherche de la distinctivité maximale dans une situation où le degré d'information apporté par le contexte est nul ? D'une certaine façon, la distinctivité du système est assurée par une hyper-articulation lorsque l'information contextuelle est absente, ou par les informations contextuelles lorsque la distinctivité articulatoire ne peut être atteinte.

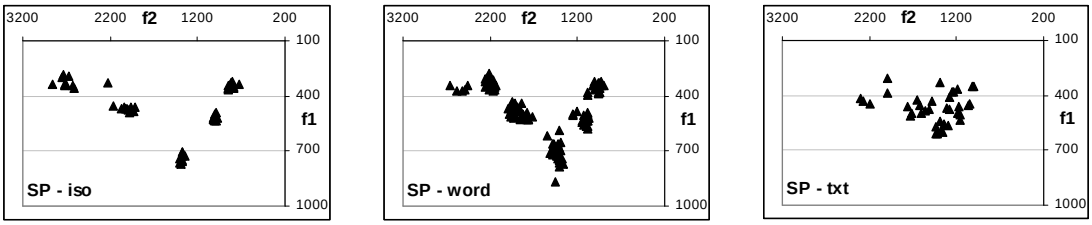


Figure 7: valeurs F1/F2 des voyelles produites par le **locuteur espagnol** LJ dans les trois corpus

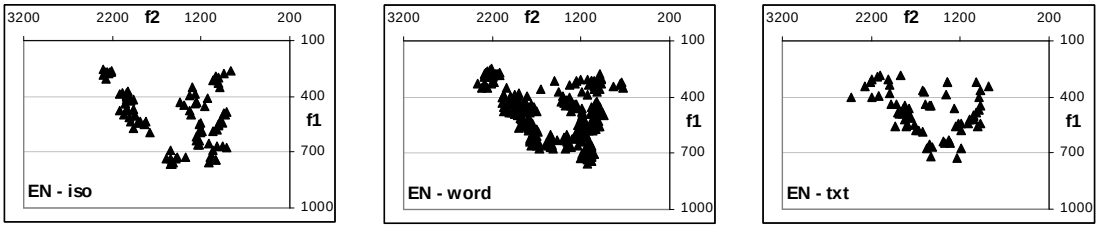


Figure 8: valeurs F1/F2 des voyelles produites par le **locuteur anglais** RN dans les trois corpus

BIBLIOGRAPHIE

[1] B. Lindblom. Phonetic Universals in Vowel Systems. In *Experimental Phonology*. Edited by Ohala J.J. & Jaeger J.J., Academic Press Inc., 13-44, 1986.

[2] J.L. Schwartz, L.J. Boë, N. Vallée. Major trends in vowel system inventories. *Journal of Phonetics*, 25, 233-253, 1997.

[3] Johnson, K. Adaptative dispersion in vowel perception. *Phonetica*, 57, 181-188, 2000.

[4] S.Y. Manuel, R.A. Krakow. Universal and Language particular aspects of vowel-to-vowel coarticulation. *Haskins Lab. S.R.S.R. SR-77/78*, 69-78, 1984.

[5] I. Maddieson, R. Wright. The vowels and consonants of Amis – a preliminary phonetic report. *UCLA Working Papers in Phonetics*, Phonetics Laboratory, Los Angeles, CA, USA, 45-65, 1995.

[6] C. Meunier, C. Frenck-Mestre, T. Lelekov-Boissard, M. Le Besnerais. Production and perception of foreign vowels: does the density of the system play a role? *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 723-726, 2003.

[7] B. Lindblom. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781, 1963.

[8] C. Gendrot & M. Adda-Decker. Analyses formantiques de corpus radiophoniques multilingues. *Actes de la conférence MIDL 2004*, Paris, Novembre 2004.

[9] B. Lindblom. Explaining phonetic variation: a sketch of the hyper- and hypospeech theory. *Speech Production and Speech Modelling*, Hardcastle and Marchal (eds.), Kluwer Academic Publishers, 403-439, 1990.

[10] N. Vallée. *Systèmes vocaliques : De la typologie aux prédictions*. Thèse de Doctorat en Sciences du Langage, Université de Grenoble, 1994.

Natures de schwa en gallo (ou « il y a schwa, schwa et schwa »)

Jean-Pierre Angoujard

Laboratoire de Linguistique de Nantes (LLING, EA 3827)
Université de Nantes - Faculté des Lettres et Sciences Humaines
Chemin de la Censive du Tertre, BP 81227, 44312 Nantes Cedex 3, France
Mél. : jean-pierre.angoujard@univ-nantes.fr

ABSTRACT

In this paper we offer a first declarative analysis of Gallo schwa (Gallo is a Romance language spoken in eastern Brittany). Whereas the behaviour of schwa in French can be derived from the properties of a unique object, the various achievements of schwa in Gallo must be related to distinct objects : a default vowel (as in French), a lexical vowel, and an optional vocoid preceding a syllabic consonant. We will show that the major properties of Gallo schwa̲s can be described through only two constraints, of which one is lexical and the other is rhythmic.

1. INTRODUCTION

Si le comportement et la nature du schwa en français peuvent être aisément rapportés aux *propriétés* d'un objet unique (section 2), les diverses réalisations de type schwa en gallo[1] paraissent bien devoir être associées à des objets distincts. À côté des alternances du type [sǝm] ∼ [sme] (pour *i sem* « il sème » *vs vou smée* « vous semez »), très largement comparables à l'alternance régulière entre [ǝ] et ∅ en français, on rencontre un ensemble de voyelles centrales (de réalisation [ǝ]) et dont la présence est constante (de type [ardǝ] *arder* « brûler », section 3). On observe également des réalisations de vocoïdes devant [r] suivi de consonne, données qui ont régulièrement été interprétées [6], par comparaison avec le français, comme le résultat d'une « métathèse » de schwa : ainsi, [vãdᵊrdi] *vandrdi* « vendredi » (section 4) .

Cet article a pour objectif de proposer une première analyse *déclarative* [3, 2] des propriétés caractéristiques des réalisations de type schwa en gallo. La comparaison entre les formes du gallo et celles du français sera ainsi rapportée (au sein d'une typologie par principes et paramètres) à seulement deux contraintes, dont l'une est lexicale et la seconde rythmique (section 5).

2. LE SCHWA EN FRANÇAIS

De nombreuses langues font usage d'une voyelle « par défaut ». Cette voyelle peut être un [ɪ], comme en arabe tunisien ; la voyelle centrale haute [ɨ], comme dans certains dialectes marocains ; un schwa [ǝ], comme en français[2]. Nous admettrons que tout schwa est, en français, la réalisation par défaut d'un noyau syllabique lexicalement « vide »[1].

Dans le cadre de la phonologie déclarative [3, 8] et en suivant [2], nous poserons que tout son peut être représenté sous la forme d'un trait [attribut : valeur]. Ce trait a pour valeur les traits SEG (dont la valeur est une liste d'éléments [11, 10]), POS (dont la valeur numérique indique s'il s'agit d'une attaque de syllabe (type *init*, avec [POS : 1]), d'un noyau (type *som*, avec [POS : 2]) ou d'une coda (type *fin*, avec [POS : 3])) et PC (qui rend compte de la valeur de sonorité). Un sommet lexicalement vide sera représenté comme un noyau avec une valeur non définie (une *variable*) pour le trait SEG :

$$[[\text{SEG} : Seg] \wedge [\text{POS} : 2] \wedge [\text{PC} : Pc]]$$

L'instanciation de la valeur du trait SEG est réalisée par la règle de défaut *e-def* ci-dessous :

$$\textit{e-def} = \text{SOM}|\text{SEG} : [1] \wedge \text{non\_var}([1]) \xrightarrow{d} [\text{ǝ}]$$

En dehors des cas d'instanciation obligatoire de la valeur du trait SEG (après une attaque syllabique de type [obtruante + liquide] ou lexicalement marquée comme dans [bǝlɔ̃] « belon »), la réalisation des schwas est essentiellement gouvernée par une contrainte rythmique [12, 1] qui exclut toute succession immédiate de noyaux « vides ». Il a été montré dans [2] que cette contrainte peut être intégrée dans la *description* du type som (la syllabe comprend un trait POIDS dont la valeur est un nombre entier et qui rend compte de sa place dans la hiérarchie rythmique de la séquence ; on admet que tout noyau appartenant à une syllabe de poids > 2 doit être instancié) :

$$\left\{ \begin{bmatrix} \text{SEG} : & Seg \\ \text{POS} : & 2 \\ \text{PC} : & Pc \end{bmatrix} \wedge \Big(\Big(Poids > 2 \Big) \Rightarrow \Big(\neg \text{ var}(Seg) \Big) \Big) \right\}$$

FIG. 1: Description du type *som*

Les réalisations éminemment variables (dialectales, individuelles) et qui autorisent, par exemple, aussi bien [pǝti] que [pti] « petit », comme [dvǝniʁ], [dǝvniʁ] ou encore [dǝvǝniʁ] « devenir », sont ainsi rapportées à des variations (partiellement) libres de l'organisation rythmique de la séquence.

[1] Le gallo est une langue d'oïl parlée en Haute-Bretagne (Ille et Vilaine, Loire Atlantique et parties est du Morbihan et des Côtes d'Armor). Dans cet article, les mots gallos sont en italiques, la traduction française entre guillemets.

[2] Par « français » on entendra ici le dialecte de l'auteur (originaire du Morbihan) et dont les caractéristiques concernant l'usage du schwa sont communes à une large partie nord de la France.

3. SCHWA ET VOYELLE CENTRALE EN GALLO

Le gallo fait également usage de représentations lexicales incluant un somme vide. En font état des alternances régulières, qu'elles soient systématiques comme dans [isəm] *i sem* « il sème » *vs* [ismi] *i smi* « il a semé », ou libres comme dans [dəvã] ∼ [dvã] « devant ». On retiendra, comme pour le français, la règle *e-def* (ci-dessus, section 2) et une description du type *som* faisant référence à la valeur du trait POIDS.

À côté de ces alternances [ə] ∼ ∅, on rencontre en gallo de nombreuses réalisations de [ə] qui ne sont sujettes à aucune alternance. Nous avons signalé qu'il en allait ainsi en français pour les noyaux situés après une attaque double (*cf.* [bʁətɔ̃] « breton »). Nous verrons que ce contexte est sans objet en gallo (section 4).

Il existe également en français de rares mots lexicalement marqués, comme [bəlɔ̃] « belon » (à comparer avec « melon ») ou [dəɔʁ] « dehors ». La présence de voyelles centrales stables[3] est, par contre, tout à fait régulière en gallo. On les retrouve notamment à l'infinitif des verbes du 1^er groupe : [dɔ̃tə] *donter* « dresser », [sərə] *sérer* « cueillir » etc. Elles sont également présentes dans de nombreux mots comme [prə] *pre* « pré », [bɔ̃tə] *bonte* « bonté » ou encore [jər] *ier* « hier »[4] .

La description des formes incluant une voyelle centrale stable est immédiate : leur réalisation est *contrainte* par leur représentation lexicale. On comparera, par exemple, les représentations lexicales des mots *bonte* (gallo) et *bonté* (français)[5] :

[bɔ̃tə] = {b, ɔ, N, t, ə} *vs* [bɔ̃te] = {b, ɔ, N, t, e}

Cette présence lexicale de [ə] en gallo se distingue très clairement de la présence régulière de sommets vides. La représentation lexicale du mot « devant » (tant en français qu'en gallo) sera tout simplement : {d, v, a, N, (t)}.

4. DES SONANTES SYLLABIQUES

4.1. Une métathèse de schwa ?

Les présentations du gallo font quasi systématiquement référence à une « interversion du *r* et de la voyelle *ė* » (p. 147 de [6], où *ė* représente la voyelle [ə]). Sont ainsi citées, en appui à cette analyse, des formes comme (la transcription est celle retenue dans [6]) :

bėrbi « brebis », *bėrtõ* « breton », *i pėrnë* « ils prenaient » , *kėrvė* « crever », *pėrswė* « pressoir », *gėrnyė* « grenier ».

Cette conception n'a pas seulement l'inconvénient majeur (*via* l'usage de termes comme « interversion » ou encore, ailleurs, « métathèse ») de laisser imaginer que le gallo serait une transformation du français, elle a avant tout pour effet de dissimuler la propriété attachée au *r* lorsqu'il est précédé et suivi d'une consonne.

Il a fallu attendre une première analyse phonologique, réalisée dans le cadre d'un mémoire de maîtrise à l'Université de Rennes 2, pour que l'hypothèse d'une *syllabification* de *r* soit retenue [4], soit, pour les mots *berton* et *gernye*, des réalisations [bṛtɔ̃] et [gṛnjə].

Le *r* syllabique ne se rencontre pas uniquement derrière une consonne (là où le français réalise une attaque double), mais également en position initiale de mot et en l'absence de voyelle adjacente. On opposera ainsi les réalisations *la rvanch* [larvãʃ] « la revanche » et *rvanch ta* [ṛvãʃta] « venge-toi ».

Il est plutôt surprenant de voir que l'existence de formes comme *prie* [pṛjə] « prier » ou encore *brouett* [bṛwɛt] « brouette » n'aient pas évité, dès l'origine, toute référence à des métathèses : la métathèse de la forme française [bʁuɛt] ne saurait pourtant être que *[buɾɛt]. . . Ajoutons que ces approches par métathèses sont naturellement diachroniques (on « explique »une forme par son origine supposée) et que le moindre souvenir, à coté de *brouett*, du mot latin *birota* (ou, pour *brbi*, de *berbice(m)*) devrait rendre sceptique. Récemment, une grammaire synchronique se rapproche davantage de la réalité phonologique en parlant de « voyelle d'appui devant un *r* » [9].

La consonne *l*, seconde liquide coronale du gallo, est également susceptible d'occuper une position syllabique. On la rencontre essentiellement, sous cette forme, en position finale de mot : *i subll* [isybl̩] « il siffle »,'*il'anfll* [ilɑ̃fl̩] « il enfle » etc.[6] La syllabification de *l* est beaucoup plus restreinte que celle de *r* pour une raison simple : on ne rencontre jamais de *l* derrière consonne en initiale de mot, mais (là où le français aurait *l*) un [j] (ici représenté orthographiquement comme *lh*) : *blhèser* [bjɛsə] « blesser », *flhourr* [fjur] « fleur », *plhée* [pje] « pluie » etc.

4.2. Syllabification et voyelle d'appui

Il est tout à fait courant de rencontrer, pour des formes jugées identiques, des représentations phonétiques faisant état d'une consonne syllabique et des représentations incluant un schwa. Ainsi, pour l'anglais, [bʌtn̩] et [bʌtən̩] « button », ou encore [bɒtl̩] et [bɒtəl̩] « bottle ». Ces représentations sont généralement rapportées à la vitesse d'élocution, la consonne syllabique étant caractéristique du débit rapide.

Ces réalisations concurrentes ont conduit à proposer des représentations phonologiques plurilinéaires dans lesquelles la consonne susceptible d'être syllabique est, dans tous les cas, associée à la coda [1]. Le noyau sera quant à lui associé à schwa ou à *r* (alors doublement associé). Cette approche a l'inconvénient de retenir une association systématique des consonnes syllabiques à deux positions rythmiques (aux position 2 et position 3) et donc une identité stricte entre, par exemple, [bər] et [bṛ] (compris comme [b*ṛ], où * représente un noyau lexicalement vide et associé au *r*).

Dans un cadre déclaratif (et donc « orienté surface »), il est intéressant de se référer à la coordination temporelle et aux chevauchements [5, 7]. Dans les schémas ci-dessous,

[3] Même si la réalisation est également de type [ə], il est sans doute préférable de réserver le terme de « schwa » aux seules voyelles caractérisées par leur alternance régulière avec ∅.

[4] Cette réalisation comme voyelle centrale est présente sur la majeure partie de la zone linguistique concernée. La réalisation est de type [e], comme en français, dans les parties littorales nord et sud [6].

[5] Toute voyelle nasale correspond à une représentation lexicale de type voyelle + élément **N**.

[6] Ces réalisations syllabiques de *l* sont caractéristiques du nord de la zone gallèse.

la durée de la voyelle demeure coextensive à celle des consonnes, seul l'écart entre les deux consonnes varie (jusqu'à zéro lorsque la consonne est syllabique) :

[bər] [bɻ]

Si la réalisation syllabique de *r* est attendue en gallo (et donc, ici, la succession immédiate des consonnes [b] et [ɻ]), un intervalle entre les deux consonnes peut être produit (d'étendue variable et associé à une variation du débit). Ceci peut être une explication des transcriptions variées rencontrées dans la littérature, soit, par exemple, [bɻtɔ̃], [bᵊrtɔ̃] ou [bərtɔ̃] pour le mot signifiant « breton »[7].

4.3. Une lecture paramétrique

Le français fait un usage régulier d' « attaques doubles » (séquences [obstruante + {ʁ ∨ l}]). Dans ces séquences, le type *init* contient deux segments consonantiques associés à une position d'attaque syllabique unique (à une position de valeur 1) et la position 2 (noyau, type *som*) est nécessairement instanciée :

$$\left(\begin{array}{l} {\scriptstyle init}\left[\begin{array}{ll}\left[\begin{array}{ll}\text{SEG}: & Seg_{\boxed{1}}\\ \text{POS}: & 1_{\boxed{2}}\\ \text{PC}: & i\end{array}\right] & \left[\begin{array}{ll}\text{SEG}: & Seg_{\boxed{2}}\\ \text{POS}: & \boxed{2}\\ \text{PC}: & p\end{array}\right]\end{array}\right] \\ {\scriptstyle som}\left[\begin{array}{ll}\text{SEG}: & Seg_{\boxed{3}}\\ \text{POS}: & 2\\ \text{PC}: & s\end{array}\right] \wedge (p > i) \wedge \neg\,(\text{var}(\boxed{3})) \end{array} \right)$$

FIG. 2: Contrainte att-double

La valeur du trait *som*|SEG peut être instancié lexicalement (par exemple comme [i] dans le mot « abri » , avec pour représentation lexicale {a, b, ʁ, i}). Si tel n'est pas le cas, la règle *e-def* (section 2) s'applique, pour conduire par exemple à [bʁətɔ̃] « breton » (la représentation lexicale est {b, ʁ, t, ɔ, N}).

Les représentations lexicales sont des suites ordonnées de segments[8]. L'organisation rythmique (syllabique) est dépendante de l'organisation en groupe phonologique (*phrase*) [2]. On représentera comme (*C_Ryth*) la contrainte qui impose que toute séquence sonore (type *phrase*) soit associée à un rythme (à une structure syllabique, *i.e.* à la succession régulière d'un nombre n de « modèles rythmiques » [1] ou types mod)[9] :

C_Ryth = ($phrase \Rightarrow mod^{+}$)

Pour toute langue qui, comme le français, fait usage de *positions vides*, *i.e.* de positions rythmiques auxquelles n'est associé aucun segment, l'interprétation rythmique ne saurait être immédiate. Ainsi, appliquées à une représentation lexicale comme {b, ʁ, t, ɔ, N}, les seules contraintes du modèle rythmique autoriseraient, pour le français, au moins deux structures (les sommets vides sont représentés par des « * ») :

a) = [[b ʁ *] [t ɔ̃]]
b) = [[b *] [ʁ *] [t ɔ̃]]

Il est alors fait référence à un principe de minimalité (*Min*), qui impose que la structuration rythmique minimale soit retenue, *i.e.* la structure comportant le moins de sommets vides. La contrainte *Min* exclut la structure b) ci-dessus.

Une troisième structure, soit ? ?[b * ʁ t ɔ̃], supposerait que la liquide [ʁ] soit associée à une position 3 (une coda). Cette interprétation conduirait à une structure rythmique comprenant un modèle rythmique à trois positions (de type *Mod_H*) et dont le noyau ne serait pas lexicalement instancié. Cette configuration est exclue par référence à la définition du type *Mod_H* :

$$Mod\_H = \Big(init \wedge som|[\text{SEG} : \boxed{1}] \wedge fin\Big) \wedge \neg(\text{var}(\boxed{1}))$$

On admettra enfin que, pour toute langue, il existe une contrainte décrivant la liste des sons susceptibles d'être associés à une position 2 (à un noyau). Cette contrainte correspond à un sous-ensemble des associations autorisées par une contrainte universelle excluant les obstruantes de cette position (toute obstruante contient l'élément **h** dans son expression) :

$$S\_pos2 = \left[\begin{array}{ll}\text{SEG}: & \boxed{1}\\ \text{POS}: & 2\end{array}\right] \wedge \neg(\text{dans(h, } \boxed{1}\text{))}$$

Nous retiendrons un paramètre lié à la contrainte S_pos2, avec pour valeur *liq_pos2 ? (oui/non)*. Pour le français, les liquides, comme les obstruantes, ne peuvent être associées à une position 2, *i.e.* il n'existe pas de liquide *syllabique* (valeur *non* pour *liq_pos2*). Si nous admettons que les liquides [ɾ] et [l] peuvent être syllabiques en gallo (valeur *oui* pour *liq_pos2*), et si nous supposons que le gallo et le français partagent une même représentation lexicale (*modulo* [ʁ]/[ɾ]) pour le mot signifiant « breton », alors trois structurations rythmiques sont envisageables en gallo et la structure c) ci-dessous sera clairement retenue par application de la contrainte *Min* :

a) = [[b ɾ *] [t ɔ̃]]
b) = [[b *] [ɾ *] [t ɔ̃]]
c) = [[b ɻ] [t ɔ̃]]

5. DÉCLARATIVITÉ, PRINCIPES ET PARAMÈTRES

Le tableau 1 fournit les valeurs retenues pour les contraintes (ici pertinentes) en français et en gallo. Lorsqu'une contrainte correspond à un principe (lorqu'elle est, par hypothèse, d'application pour toute langue), celle-ci a la valeur **U** (pour « universelle »). Les valeurs paramétriques sont notées **O**(ui) ou **N**(on). La contrainte *N_pos2* conduit à la réalisation de voyelles nasales, soit $< \text{V, N} > \equiv \tilde{\text{V}}$.

Dans la mesure où, pour la phonologie déclarative, toute

[7] Cette hésitation se retrouve dans les propositions pour une orthographe unifiée du gallo, où se sont rencontrées récemment les formes *brton*, *b'rton* ou *berton*

[8] Il s'agit là d'une simplification dans la mesure où les segments correspondent à des *expressions* (ou combinaisons d'éléments) [11, 10].

[9] Le constituant appelé « modèle rythmique » correspond à la succession de 2 ou 3 positions rythmiques, soit [pos 1 ℘ pos 2 (℘ pos 3)].

	français	gallo
C_Ryth	**U**	**U**
Mod_H	**U**	**U**
Min	**U**	**U**
Att_double	**O**	**O**
N_pos2	**O**	**O**
e_lex	**N**	**O**
liq_pos2	**N**	**0**

TAB. 1: Contraintes

contrainte est identifiée à une *représentation partielle*, les deux réalisations [bʁətɔ̃] (pour le français) et [bɾ̥tɔ̃] (pour le gallo) peuvent être respectivement *identifiées* aux conjonctions de contraintes (non ordonnées) ci-dessous :

[bʁətɔ̃] ≡ {*C_Ryth* ∧ *Mod_H* ∧ *Min* ∧ *N_pos2* ∧ *Att_double* ∧ <b, ʁ, t, ɔ, N>}

[bɾ̥tɔ̃] ≡ {*C_Ryth* ∧ *Mod_H* ∧ *Min* ∧ *N_pos2* ∧ *liq_pos2* ∧ <b, ʁ, t, ɔ, N>}

Remarque : La contrainte *Att_double* est également active dans la phonologie du gallo (voir, par exemple, [bru] *brout* « lierre », [frãʒə] *franjer* « déchirer » etc.). Simplement, en l'absence de voyelle lexicale, la représentation rythmique minimale est retenue : {*Min* ∧ *liq_pos2*}.

On constatera également que cette analyse par contraintes rend compte, sans le moindre ajout, des oppositions français *vs* gallo du type [bʁuɛt] *vs* [bɾ̥wɛt] et pour lesquelles aucune référence à une quelconque « métathèse » n'est imaginable. Il suffit de faire l'hypothèse que ces formes ont une représentation lexicale identique dans les deux langues (*modulo* les réalisations spécifiques [r/ʁ]), soit {b, r/ʁ, U, ɛ, t} :

⇒ En français, seul l'élément **U** peut être associé à une position 2 (et réalisé comme [u]) ou à une position 1 (et réalisé comme [w]). L'interprétation rythmique imposée par *Min* sera donc [[bʁu][ɛt]] (et non *[[bʁ@][wɛt]])[10].

⇒ En gallo, l'élément **U** et le segment [r] peuvent être associés à une position 2 (comme [u] et [ɾ̥]) ou à une position 1 (comme [w] et [r]). La réalisation syllabique de la liquide conduit à la structuration rythmique minimale [[bɾ̥][wɛt]].

6. CONCLUSION

Si le gallo partage avec le français l'usage d'une voyelle de type [ə], il s'en distingue crucialement dans la mesure où cette voyelle correspond, en gallo, à deux *objets* dont les *propriétés* sont clairement distinctes :

- Une voyelle appartenant au système vocalique du gallo et présente lexicalement. On opposera ainsi les représentations lexicales {g, ə, ɲ, ə} (gallo) et {g, a, ɲ, e} (français) (verbe « gagner »).
- Une voyelle qui n'est rien d'autre que la réalisation par défaut d'un sommet vide, réalisation gouvernée par l'organisation rythmique. Ces alternances entre ə et ∅ se retrouvent en gallo comme en français (mais également dans toute langue qui fait usage de sommets vides et de voyelle par défaut, comme par exemple, *modulo* la qualité de la voyelle, les dialectes arabes du Maghreb). L'organisation rythmique de la langue considérée rendra compte du fait que tel ou tel sommet vide sera associé à une position rythmique « forte » et prioritairement instancié (pour le français, voir [12, 1] ; pour le gallo, cette description reste à accomplir).

À ces deux types de [ə], il faut ajouter les réalisations occasionnelles d'un vocoïde devant *r* ou *l*, réalisations qui doivent être interprétées comme des variantes des réalisations syllabiques [ɾ̥] et [l̥] (valeur *oui* pour le paramètre *liq_pos2*).

L'analyse déclarative présentée, qui ne suppose donc aucune dérivation et qui se réfère à un niveau unique de représentation, permet d'opposer les réalisations gallèses et françaises par la seule présence et/ou absence d'une ou plusieurs contraintes (ou représentations partielles) au sein de la forme attestée (ou représentation), forme en tout point équivalente à la liste non ordonnées des contraintes qui la définissent.

[10] Nous n'avons pas d'explication au fait que la structuration *[bʁwɛt], assurément « minimale », soit exclue ; on peut seulement noter que l'unique voyelle attestée derrière [Cʁw] est [a].

RÉFÉRENCES

[1] Jean-Pierre Angoujard. *Théorie de la Syllabe. Rythme et Qualité*. CNRS Editions, Paris, 1997. 224 p.

[2] Jean-Pierre Angoujard. *Phonologie Déclarative*. CNRS Editions, Paris, 2006. 198 p.

[3] Steven Bird. *Computational Phonology. A constraint-based approach*. Cambridge University Press, Cambridge, 1995. 203 p.

[4] Bertran Ôbrée. Les sonantes et la syllabe en gallo, 1998. Mémoire de maîtrise, Université de Rennes 2 Haute-Bretagne.

[5] Catherine P. Browman and Louis Goldstein. Tiers in articulatory phonology, with some implications for casual speech. In John Kingston and Mary E. Beckman, editors, *Papers in Laboratory Phonology I*, pages 341–376. Cambridge University Press, Cambridge, 1990.

[6] Jean-Paul Chauveau. *Le gallo : une présentation*. Université de Bretagne Occidentale, Brest, 1984. 252 p.

[7] John Coleman. The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology*, 9(1) :1–44, 1992.

[8] John Coleman. *Phonological representations. Their names, forms and powers*. Cambridge University Press, Cambridge, 1998. 345 p.

[9] P. Deriano. *Grammaire du gallo*. Éditions Label LN, Ploudalmézeau, 2005.

[10] John Harris. *English Sound Structure*. Blackwell, Oxford, 1994. 317 p.

[11] Jonathan Kaye, Jean Lowenstamm, and Jean-Roger Vergnaud. The internal structure of phonological elements : a theory of charm and government. *Phonology*, 2 :303–326, 1985.

[12] Elizabeth Selkirk. The French foot : on the statute of "mute" *e*. *Studies in French Linguistics*, I-2 :79–141, 1978.

Dimensions acoustiques de la parole expressive : poids relatifs des paramètres resynthétisés par Praat vs. LF-ARX

Nicolas Audibert[1], Damien Vincent[2], Véronique Aubergé[1], Albert Rilliard[1] & Olivier Rosec[2]

[1] Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
[2] France Telecom, R&D Division
{audibert, auberge, rillard}@icp.inpg.fr, {damien.vincent, olivier.rosec}@francetelecom.com

ABSTRACT

The emotional prosody is multi-dimensional. A debated question is whether some parameters are more specialized to convey some emotion dimensions. Selected stimuli carrying acted expressions of anxiety, disappointment, disgust, disquiet, joy, resignation, satisfaction and sadness on monosyllabic words were used to synthesize artefactual stimuli by projecting separately prosodic parameters on neutral expressions, with Praat and an LF-ARX algorithm. Perceptive evaluation of stimuli and comparison of results (1) indicate that F0 contours bring more information on positive expressions while voice quality and duration convey more information on negative expressions, and intensity alone is not informative enough (2) diagnoses minor artifacts of both synthesis methods which consequences may have interesting implications in expressive speech synthesis (3) validates the efficiency of the LF-ARX algorithm (4) measures the relative weights of each of the LF-ARX voice quality parameters.

1. INTRODUCTION

La multi-dimensionnalité de la prosodie des affects est un problème complexe, le débat restant ouvert quant à la spécificité de certains indices émotionnels pour l'expression d'affects particuliers. Une autre question non résolue est celle de la description de la qualité de voix comme une seule ou plusieurs dimensions. Afin d'évaluer le poids de chaque paramètre prosodique dans la morphologie vocale des émotions, et donc dans la perception de leurs expressions, il est nécessaire d'évaluer comment des stimuli porteurs de ces expressions sont perçus lorsqu'on agit séparément sur ces paramètres. Étant donné que les variations de ces paramètres résultent d'un contrôle global du conduit vocal, il semble toutefois impossible de recueillir naturellement de tels stimuli, quand bien même des locuteurs seraient spécifiquement entraînés pour cette tâche.

Nous avons donc adopté une méthode basée sur la resynthèse : une analyse acoustique de stimuli de référence est effectuée avant de synthétiser de nouveaux stimuli à partir de tout ou partie des paramètres analysés, en fonction des hypothèses à tester ; une évaluation perceptive de ces stimuli permet ensuite de valider ces hypothèses. Une telle méthode a déjà été utilisée dans une série d'études relatives à l'expression des affects dans la parole. Ainsi, le rôle de la qualité de voix dans les expressions émotionnelles a été attesté à l'aide de stimuli resynthétisés après modification de l'onde de débit glottique analysée sur des stimuli de référence [8]. Par ailleurs [5], les variations de F0 et de durée extraites de diverses expressions émotionnelles ont été appliquées en synthèse par concaténation à des diphones porteurs d'expressions émotionnelles différentes. L'évaluation perceptive des stimuli ainsi construits a conduit les auteurs à conclure que les expressions de colère étaient majoritairement véhiculées par les diphones, la tristesse par F0 et la durée, tandis qu'aucune conclusion claire ne pouvait être tirée quant aux expressions de joie. Enfin, dans un certain nombre d'études (par ex. [3]), des stimuli ont été synthétisés à partir de mesures multiparamétriques afin d'évaluer la pertinence des paramètres acoustiques mesurés pour la perception d'expressions émotionnelles.

Deux méthodes ont été successivement utilisées pour réaliser cette resynthèse paramètre par paramètre, les stimuli générés étant ensuite évalués perceptivement selon le même protocole. La première (présentée en section 2), basée sur Praat [4], ne permet pas de manipuler directement la qualité de voix. En revanche la seconde (présentée en section 3), basée sur un modèle ARX [6] excité par une source LF [7], permet d'évaluer séparément le rôle des différents paramètres de qualité de voix.

2. ETUDE 1 : RESYNTHÈSE PRAAT

2.1. Choix des données de référence

10 stimuli extraits du corpus E-Wiz / Sound Teacher [1] et constituant un sous-ensemble des 72 stimuli actés précédemment évalués dans une étude perceptive [10] ont été sélectionnés. Ces stimuli expriment sur les mots monosyllabiques [ʁuʒ] et [sabl] les états émotionnels suivants, joués par un acteur après avoir été ressentis dans une tâche « piège » prétextant une aide à l'apprentissage phonétique des langues étrangères : anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction, tristesse et neutre. La pertinence d'une expression neutre, sélectionnée comme référence pour la comparaison des contours mutiparamétriques, est validée par la présence dans les productions spontanées de ce locuteur d'expressions étiquetées par lui-même comme « rien ».

2.2. Méthode de synthèse

L'analyse des contours multiparamétriques et la synthèse ont été effectuées à l'aide la fonction de synthèse basée sur TD-PSOLA de Praat [4], selon une procédure semi-

automatique consistant à styliser les contours de F0 et d'intensité extraits d'un stimulus source puis à appliquer tout ou partie de ces contours à un stimulus cible et générer un nouveau stimulus combinant des propriétés des stimuli source et cible. La stylisation et la transplantation des contours ont été contrôlées afin d'éviter de négliger des points saillants. La transplantation des contours d'intensité a été réalisée en appliquant un contour relatif puis en rééchelonnant le signal pour obtenir la même valeur globale d'énergie.

2.3. Stimuli générés

Pour chacun des stimuli originaux porteurs d'une expression émotionnelle, 5 stimuli distincts ont été générés, étiquetés en fonction des paramètres du stimulus cible utilisés : (i) *contrôle*, construit en appliquant les contours stylisés de F0 et d'intensité du stimulus source à lui-même et destiné à évaluer d'éventuels artefacts dus au processus de resynthèse (ii) *F0*, construit en appliquant le contour stylisé de F0 du stimulus source au stimulus porteur d'une expression neutre correspondant au même mot (iii) *intensité,* obtenu en appliquant le contour d'intensité du stimulus source à l'expression neutre correspondante (iv) *F0 et intensité,* construit en appliquant les contours de F0 et d'intensité à l'expression neutre correspondante (v) *QV et durée*. Cette dernière condition a été obtenue en appliquant les contours de F0 et d'intensité de l'expression neutre au stimulus source. Ainsi seuls les phénomènes de durée et de qualité de voix (QV) du stimulus source subsistent, tandis que ses variations spécifiques de F0 et d'intensité sont neutralisées. En complément des 40 stimuli générés à partir des 8 expressions émotionnelles sélectionnées, un stimulus en condition *resynthèse complète* a été généré pour chacune des 2 expressions neutres.

2.4. Evaluation perceptive

Les 42 stimuli générés ont été notés par 40 juges de langue maternelle française (6 hommes, 34 femmes, d'âge moyen 23,3 ans) en chambre sourde, en ordre aléatoire, avec 3 présentations non consécutives de chaque stimulus La présentation des stimuli et l'enregistrement des réponses ont été automatisés à l'aide d'une interface. Les sujets avaient pour instruction de sélectionner l'une des 8 expressions proposées (anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction ou tristesse) ou l'étiquette neutre. De plus il leur était demandé de noter l'intensité émotionnelle perçue entre 1 et 10.

2.5. Résultats

La valeur élevée de l'alpha de Cronbach (α=.95) indique que les réponses données par les différents juges sont cohérentes. Les résultats ont été traités séparément pour chaque condition de resynthèse. Etant donné que les intensités émotionnelles sont significativement corrélées au nombre de réponses attribuées aux différentes étiquettes (r^2=.854) et apportent peu d'information supplémentaire, seuls les résultats relatifs aux scores d'identification sont discutés ici. Afin de prendre en compte les principales confusions faites par les sujets en condition de contrôle, similaires à celles observées lors de la validation perceptive des stimuli originaux [10], et faire ressortir plus clairement les principales tendances, certaines étiquettes ont été regroupées : joie avec satisfaction, anxiété avec inquiétude, tristesse avec déception et résignation, tandis que dégoût et neutre demeurent des catégories distinctes. Le test du khi-deux indique que les distributions après regroupements sont significativement différentes du hasard (p=.01). Une fois ces regroupements effectués, l'essentiel des informations pertinentes se trouve sur les diagonales des matrices de confusion, qui correspondent aux scores d'identification. Les données après regroupement ont donc été converties en bonnes ou mauvaises réponses et normalisées pour permettre une évaluation statistique des différences observées au moyen d'une ANOVA sur mesures répétées.

Une première observation est que les scores obtenus en conditions « manipulées » sont moins élevés qu'en condition de contrôle, indiquant qu'aucune dimension ne véhicule seule toute l'information affective. Néanmoins cette différence n'est pas significative en condition *QV et durée* pour anxiété et inquiétude (47,2% vs. 55,6% en condition de contrôle) ni pour tristesse, résignation et déception (55,8% vs. 59,6%), la qualité de voix et la durée apparaissant comme véhiculant l'essentiel de l'information pour ces expressions négatives. Bien que ce soit également le cas pour le dégoût, les autres dimensions ne portant que très peu d'information, il est surprenant de constater que cette expression est significativement moins bien reconnue qu'en condition de contrôle (34,2% vs. 61,7%). En condition *intensité*, seules les expressions de tristesse, déception et résignation (identifiées à 35,6%) ont obtenu un score supérieur à 10%. Toutefois de nombreux sujets percevant l'une de ces expressions en condition *intensité* ont également confondu les expressions neutres avec tristesse, résignation ou déception en condition de contrôle. Ceci explique également le score plus élevé en condition *F0* (36,9%) qu'en condition *F0 & intensité* (26,7%) obtenu par ces expressions. En conditions *F0* et *F0 et intensité*, les expressions de joie et satisfaction ont été les mieux reconnues (respectivement à 56,3% et 67,5% vs. 85,4% en condition de contrôle). De plus, bien que l'intensité soit insuffisante seule, la comparaison des scores en conditions *F0* et *F0 et intensité* montre qu'elle apporte un gain significatif (p=.05) pour les expressions de joie et satisfaction (56,3% vs. 67,5%) ainsi que d'anxiété et d'inquiétude (26,7% vs. 21,3%), et qu'elle ne peut donc pas être considérée comme non informative.

3. ETUDE 2 : RESYNTHÈSE LF-ARX

3.1. Méthode de synthèse

Le modèle LF de production de la parole [7], sur lequel la méthode de synthèse utilisée ici est basée, s'inscrit dans une approche source-filtre : la source correspond à l'excitation glottique modifiée pour y intégrer l'effet des lèvres (dérivateur), ou onde de débit glottique dérivée, et le filtre modélisé aux résonances du conduit vocal. Dans le cadre de la modélisation d'un son voisé par un modèle ARX [6]

excité par une source LF, l'analyse revient à estimer les 3 paramètres du modèle LF décrivant la source, une composante stochastique appelée résidu, les coefficients du filtre correspondant au conduit vocal, la fréquence fondamentale et l'énergie. Etant donné que les paramètres du filtre et du résidu peuvent être estimés par la méthode des moindres carrés une fois les paramètres LF connus, une méthode efficace basée sur une recherche exhaustive dans un espace d'ondes LF quantifiées a été proposée pour l'estimation de ces paramètres [11].

Dans ce cadre, le processus de resynthèse consiste à remplacer certains paramètres issus de l'analyse du stimulus neutre (source) par les valeurs de ces paramètres obtenues par l'analyse du stimulus « émotionnel » (cible). Ce processus comprend une procédure d'alignement ainsi qu'un algorithme de synthèse. L'alignement nécessite que les stimuli source et cible aient le même contenu phonétique : après avoir apparié les frontières phonémiques des 2 stimuli, les points analysés dans un même phonème sont reliés par un mécanisme d'interpolation linéaire. Le résultat de cet alignement doit être contrôlé car des erreurs peuvent apparaître en cas de non congruence entre les informations de voisement des stimuli source et cible. L'algorithme de synthèse, similaire à ceux utilisés pour les modifications prosodiques basées sur TD-PSOLA [9], détermine les instants de synthèse et génère pour chacun de ces instants une paire de trames issues respectivement des stimuli source et cible, la suite du processus de synthèse devenant alors triviale.

3.2. Stimuli générés

Les 10 stimuli de référence de l'étude 1 ont été à nouveau utilisés comme base pour la resynthèse par l'algorithme LF-ARX. Toutefois la méthode de synthèse utilisée n'a pas permis de générer des stimuli de qualité suffisante à partir de l'expression de la satisfaction, qui a donc dû être éliminée de cet ensemble. Seules 7 conditions de synthèse ont été retenues parmi les combinaisons possibles des 6 jeux de paramètres (la qualité de voix étant considérée comme décrite par l'ensemble source, résidu et filtre), étiquetés en fonction des paramètres du stimulus expressif utilisés : (i) *contrôle* (ii) *QV et durée* (iii) *QV* (iv) *source et résidu* (v) *source*, (vi) *durée* et (vii) *F0 et intensité*. Les conditions *contrôle*, *QV et durée* et *F0 et intensité* permettent une comparaison directe avec les résultats de l'étude 1.

3.3. Evaluation perceptive et résultats

Les 51 stimuli générés ont été évalués par 25 juges de langue maternelle française (7 hommes, 18 femmes, d'âge moyen 25,7 ans), selon le même protocole que celui utilisé dans l'étude 1.

Les résultats de cette évaluation perceptive présentent également une valeur élevée d'alpha de Cronbach (α=.92), ainsi qu'une corrélation significative entre les intensités émotionnelles perçues et le nombre de réponses attribuées aux étiquettes correspondantes (r^2=.889). De plus les confusions observées en condition de contrôle étant similaires à celles de l'étude 1, les mêmes regroupements ont été appliqués. L'expression de la satisfaction n'étant pas présente dans les stimuli utilisés ici, l'expression de la joie a été traitée comme une catégorie distincte. De même que dans l'étude 1, les données ont été converties en bonnes ou mauvaises réponses et normalisées pour tester la significativité (p=.01) des différences observées par des analyses de variance sur mesures répétées.

Les principaux résultats de l'étude 1 sont ici confirmés : la joie est significativement mieux reconnue que les autres expressions en condition *F0 et intensité*, quoique avec un score significativement moins élevé qu'en condition de contrôle (58,7% vs. 77,3%) ; les expressions de tristesse, résignation et déception, de même que celles d'anxiété et d'inquiétude ne sont pas significativement moins bien reconnues en condition *QV et durée* qu'en condition de contrôle (respectivement 52,9% vs. 56% et 60% vs. 67,3%). On retrouve également pour l'expression du dégoût le même phénomène que dans l'étude 1, cette expression étant significativement moins bien reconnue en condition *QV & durée* qu'en condition de contrôle (42,7% vs. 70,7%), alors que F0 et intensité portent très peu d'information (1,3%). Des observations peuvent en outre être tirées des conditions de synthèse absentes de la première évaluation, les dimensions testées conjointement sous l'étiquette *QV & durée* ayant ici fait l'objet d'une évaluation séparée. Ainsi en condition *durée* les expressions de tristesse, résignation et déception ont été aussi bien reconnues qu'en condition de contrôle (56,9% vs. 56%), tandis que ces expressions ont été significativement moins bien reconnues en condition *QV*, mais néanmoins au dessus du hasard (44,4%). Les expressions d'anxiété et d'inquiétude, quant à elles, présentent la même quantité d'information affective en condition *QV* et en condition *durée* (identifiées à 46% dans ces 2 conditions). Enfin, la majeure partie de l'information affective pour le dégoût est portée par la durée (49,3% vs. 8% en condition *QV* et 1,3% en condition *F0 et intensité*). D'autre part la comparaison des scores obtenus en conditions *VQ*, *source & résidu* et *source* permet d'évaluer les influences relatives des différents paramètres de modélisation de la qualité de voix. Si la source apparaît comme porteuse de toute l'information de qualité de voix pour les expressions de tristesse, résignation et déception ainsi que pour la joie (pas de gain significatif en ajoutant le filtre et le résidu), les expressions d'anxiété et d'inquiétude ont été significativement mieux reconnues en condition *source & résidu* (35,3%) qu'en condition *source* (24%), et significativement mieux en condition *QV* (46%) qu'en condition *source & résidu*.

Etant donné les scores obtenus pour l'expression de la joie en condition *QV* (10,7%) et en condition *durée* (0%), on devrait également obtenir un score d'environ 10% pour cette expression en condition *QV & durée*. Or ce score est largement supérieur à la valeur attendue (30,7%), ce qui nous a alerté sur la possible présence d'un artefact. Un examen attentif des signaux a révélé la présence d'un bruit de fermeture très court et d'énergie moyenne au début du signal généré en condition *QV & durée*, qui peut avoir été

interprété par les juges comme un coup de glotte annonciateur d'un rire, rendant le score d'identification correspondant artificiellement élevé. En effet l'algorithme de synthèse, étalonné sur des signaux dans lesquels les portions étiquetées comme silence sont effectivement silencieuses, génère les segments silencieux en copiant les parties étiquetés comme silence du stimulus source ou cible. Ce bruit étant présent dans l'expression neutre utilisée, il a donc été automatiquement copié au début du signal généré dans cette condition.

4. DISCUSSION

La table 1 présente les scores obtenus après regroupement dans les études 1 et 2 (étiquetés respectivement *Praat* et *ARX*) pour les conditions de synthèse communes. Afin de rendre les comparaisons possibles, les confusions des expressions de joie avec la satisfaction ont ici été prises en compte dans le calcul des scores d'identification de la joie dans l'étude 1, d'où la différence entre certains scores présentés dans cette table et ceux présentés en section 2.

Table 1 : Scores d'identification obtenus dans les études 1 et 2 après regroupement.

		contrôle	F0+int	QV+dur
joie	Praat	70,8%	42,5%	6,7%
	ARX	77,3%	58,7%	30,7%
trist, res, décep	Praat	59,6%	26,7%	55,8%
	ARX	56%	27,1%	52,9%
anxiété, inq,	Praat	55,6%	21,4%	47,2%
	ARX	67,3%	40,7%	60%
dégoût	Praat	61,7%	3,3%	34,2%
	ARX	70,7%	1,3%	42,7%
neutre	Praat	31,7%		
	ARX	52,7%		

Les scores obtenus en condition de contrôle sont généralement plus élevés avec la synthèse LF-ARX qu'avec la synthèse Praat, à l'exception des expressions de tristesse, résignation et déception pour lesquelles cette différence est faible (la structure des données ne permet pas de tester la significativité des différences entre les résultats des études 1 et 2). Le codage des expressions par l'algorithme LF-ARX semble donc globalement meilleur qu'avec Praat.

Par ailleurs, si on considère qu'en l'absence de l'artefact décrit en section 3 le score de 30,7% obtenu par l'expression de la joie en condition *QV & durée* dans l'étude 2 aurait dû être proche de 10%, on peut s'interroger sur la différence entre ce score théorique et celui obtenu dans l'étude 1 (6,7%). Un artefact de la synthèse Praat pourrait être responsable de ce score plus faible : la méthode de transplantation des contours d'intensité contrôlant la valeur globale d'énergie mais non les valeurs locales, l'intensité à la fin du stimulus généré est plus faible que celle à la fin de l'expression neutre. Cette intensité finale peu élevée peut donc avoir été interprétée par les juges comme incompatible avec une expression de joie, ce qui poserait alors la question de la définition d'un contour d'intensité.

Les résultats de ces deux études montrent donc que l'information affective des expressions positives est principalement véhiculée par les contours de F0, tandis qu'elle est surtout portée par la durée, et dans une moindre mesure par la qualité de voix pour les expressions négatives. Ils valident également la qualité du codage réalisé par l'algorithme LF-ARX, de même que la pertinence de la modélisation des informations de filtre et de résidu. Les conséquences des artefacts observés ont d'intéressantes implications potentielles en synthèse, puisqu'un phénomène local et aussi mineur en termes de quantité d'information peut modifier la perception des affects exprimés.

BIBLIOGRAPHIE

[1] V. Aubergé, N. Audibert and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, Lisbonne, pages 179-182, 2004.

[2] V. Aubergé, N. Audibert and A. Rilliard. Acoustic Morphology of Expressive Speech: What about Contours? *Speech Prosody 2004*, Nara, pages 201-204, 2004.

[3] T. Bänziger, M. Morel and K. R. Scherer. Is there an emotion signature in intonational patterns? And can it be used in synthesis? *Eurospeech 2003*, Genève, pages 1641-1644, 2003.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer. http://www.praat.org.

[5] M. Bulut, S. Narayanan and A. Syrdal. Expressive speech synthesis using a concatenative synthesizer. *7th ICSLP*, Denver, pages 1265–1268, 2002.

[6] W. Ding, H. Kasuya and S. Adachi. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. in *IEICE Trans. Inf. Syst.*, E78-D (6), pages 738–743, 1995.

[7] G. Fant, J. Liljencrants and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR* (4), pages 1–13, 1985.

[8] C. Gobl and A. Ní Chasaide. The role of the voice quality in communicating emotions, mood and attitude *Speech Comm.* (40), pages 189–212, 2003.

[9] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modifications of speech. *Speech Comm.* (16), pages 175–205, 1995.

[10] A. Rilliard, V. Aubergé and N. Audibert. Evaluating an Authentic Audio-Visual Expressive Speech Corpus. *4th LREC*, Lisbonne, pages 175-178, 2004.

[11] D. Vincent, O. Rosec and T. Chonavel. Estimation of LF glottal source parameters based on arx model. *Interspeech 2005*, Lisbonne, pages 333-336, 2005.

Vers un système multilinéaire de transcription des variations intonatives

Brechtje Post[1] *& Elisabeth Delais-Roussarie*[2]

[1]Research Centre for English and Applied Linguistics
University of Cambridge, UK
[2]CNRS, UMR 7110 / LLF (Laboratoire de Linguistique formelle)
Université de Paris 7, France
bmbp2@cam.ac.uk et elisabeth.roussarie@wanadoo.fr

ABSTRACT

In the paper, we will present a transcription system for Intonational Variation (IVTS), derived from IViE. The prosodic features are transcribed on i) the rhythmic tier ; ii) the local phonetic tier ; iii) the global phonetic tier ; and iv) the phonological tier. Each tier offers a range of labels which share a general architecture, but language-specific parameters determine which subset of labels a transcriber can choose from for the transcription of a particular language variety. In this paper, we will argue that the multi-linear architecture of IViE-based systems offers transparency, flexibility and standardization, three key advantages in qualitative and quantitative studies of intonational variation across languages and language varieties.

1. INTRODUCTION

La difficulté des recherches en prosodie s'explique par le fait que i) les phénomènes comme l'accentuation ou l'intonation sont par nature continus, et donc difficiles à représenter de façon discrète ; ii) les caractéristiques prosodiques varient en fonction d'éléments comme le débit ou l'origine socio-géographique des locuteurs. Aussi tout système de transcription qui permet d'encoder les événements prosodiques de façon discrète tout en prenant en compte la variation est intéressant dans de nombreux domaines : la linguistique, l'apprentissage des langues secondes, le traitement automatique de la parole, etc.

Bien que les travaux sur corpus se soient considérablement développés ces dernières années tant en linguistique qu'en traitement automatique de la parole, il n'existe pas à ce jour de système de transcription prosodique standardisé qui soit suffisamment flexible pour traiter la variation et encoder les phénomènes prosodiques dans des dialectes non encore décrits. Même si des systèmes de transcription connus et fréquemment utilisés comme INTSINT [1] ou ToBI [2] partagent de nombreuses caractéristiques avec des systèmes dérivés de IViE [3] (*Intonational Variation in English*), ils ne sont pas conçus pour travailler dans une perspective comparative et variationiste, et cela pour plusieurs raisons : i) INTSINT, en partant d'une analyse exclusivement acoustique, ne tient pas compte de données perceptives et linguistiques ; ii) ToBI, de son côté, présuppose que l'inventaire phonologique des phénomènes intonatifs soit établi puisque le transcripteur encode directement les informations au niveau phonologique ; iii) INTSINT et ToBI privilégient nettement les phénomènes intonatifs au dépens des phénomènes métriques, bien que les deux systèmes soient développés dans le cadre métrique-autosegmental ; iv) ToBI, et dans une moindre mesure INTSINT, représente le niveau local de la phrase ou de l'énoncé plutôt que le niveau global du discours.

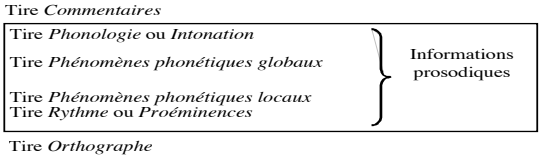
Dans ce papier, nous présentons une première version d'un système de transcription prosodique appelé IVTS (*Intonational Varitaion Transcription System*). Ce dernier est développé dans le cadre métrique autosegmental (cf. [4] et [5]) et repose sur l'idée que pour analyser adéquatement les phénomènes intonatifs, il est préférable de clairement distinguer le niveau phonologique du niveau phonétique. IVTS doit permettre d'encoder les variations intonatives afin de comparer des langues, mais aussi différents dialectes d'une même langue. Il repose sur une adaptation directe d'IViE (cf. [3], mais aussi [6] pour une application à l'allemand) qui se justifie par le fait que les systèmes comme IViE offrent de nombreux avantages en comparaison de systèmes comme ToBI ou INTSINT : i) étant conçus pour analyser les variations dialectales, ils ne présupposent pas de connaissances sur le statut phonologique de tel ou tel événement prosodique (statut métrique des syllabes proéminentes, identité phonologique d'un mouvement mélodique, etc.) ; ii) ils peuvent facilement être adaptés pour permettre l'encodage de phénomènes non locaux comme les changements de registre ou les downsteps ; iii) du fait de leur structure multilinéaire (l'encodage d'information phonétique, métrique et phonologique se faisant sur plusieurs tires ou niveaux), ils offrent plus de transparence et de flexibilité (cf. section 2 ci-après) ; iv) ils se montrent très robustes comme en témoignent les travaux menés dans le cadre d'IViE sur les proéminences (cf. [7]).

Nous allons montrer qu'un système comme IVTS permet d'avancer vers plus de flexibilité, de transparence et de standardisation, ce qui est essentiel pour mener des études prosodiques dans une perspective comparative et variationiste. Cette idée sera illustrée par des exemples précis extraits de données variées sur lesquelles nous travaillons actuellement : quatre dialectes du français (cf. [8]) et des données d'acquisition (anglais langue seconde). Pour finir, il est important de noter que ce système a vocation à être adapté et utilisé pour étudier plusieurs langues ou variétés d'une même langue, même si, pour le moment, il n'a été testé que sur un petit échantillon de données du français.

2. UN SYSTÈME MULTILINÉAIRE

Le système de transcription IVTS encode différentes informations prosodiques et orthographiques sur six niveaux (ou *tires*) distinct(e)s, parmi lesquel(le)s quatre sont spécifiquement consacré(e)s à l'annotation prosodique.

(1) les six niveaux de transcription retenus dans IVTS



2.1 Un exemple d'annotation en français

La figure 1 représente une utilisation d'IVTS pour transcrire des données du français extraites du corpus PFC. Il s'agit de l'énoncé *le village de Beaulieu est en grand émoi* lu par un homme originaire de la région de Liège (Belgique). Les six niveaux d'annotation mentionnés sous (1) sont alignés temporellement avec le signal et la courbe de fréquence fondamentale. Cette transcription a été effectuée sous Praat, mais elle pourrait l'être avec tout autre logiciel d'analyse et de visualisation du signal de parole.

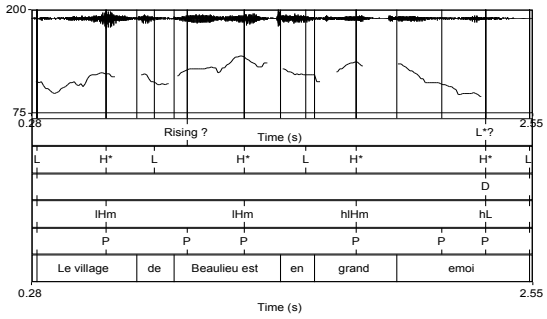


Figure 1 : Exemple d'annotation avec IVTS

Dans cette figure, les tires ou niveaux sont organisé(e)s comme suit : la première ligne correspond à la tire *orthographe*, la seconde à la tire *Rythme* (ou *Proéminences*), la troisième à la tire *Phénomènes phonétiques locaux*, la quatrième au niveau *Phénomènes phonétiques globaux* et la dernière sert à ajouter des *commentaires*.

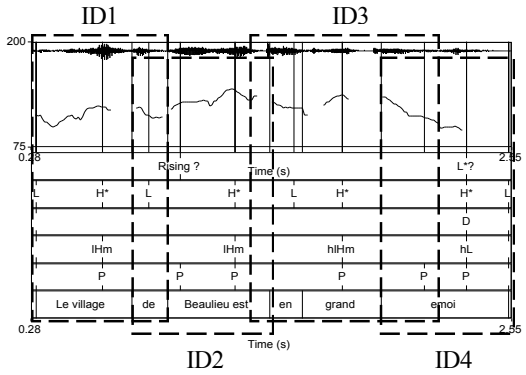
Comme on le voit, les différents mots produits sont alignés avec les portions de signal qui leur correspondent sur la tire *orthographe*. Dans les autres tires, les étiquettes sont alignées avec des points déterminés sur le signal comme : i) le milieu d'une syllabe perçue comme proéminente ; ii) les frontières de domaines intonatifs, etc. Sur la tire *Commentaires*, les points retenus pour aligner les commentaires au signal correspondent aux zones sur lesquelles portent les commentaires. Les différents alignements retenus le sont en fonction de leur pertinence dans la langue traitée et de la façon dont s'opère l'ancrage des phénomènes mélodiques.

Sur la tire *Rythme*, l'étiquette **P** indique que la syllabe marquée est plus proéminente que les syllabes adjacentes (cf. fig. 1). Cela peut se caractériser au niveau acoustique par un allongement de la durée, un mouvement mélodique, etc. (cf. [6]). Notons que **P** signale une saillance perceptive, mais pas nécessairement une propriété structurelle abstraite du mot ou du groupe de mots comme un accent lexical. De plus, si les étiquettes **P** affectent des syllabes sur lesquelles des mouvements mélodiques particuliers se réalisent, elles vont être alignées avec des étiquettes sur les tires *Phénomènes phonétiques locaux* et *Phonologie*. Ceci étant, cela n'est pas une nécessité comme on peut le voir dans la fig. 1 avec les étiquettes **P** associées aux syllabes initiales des mots *Beaulieu* et *émoi*.

La tire *Phénomènes phonétiques locaux* est utilisée pour transcrire la forme des mouvements mélodiques réalisés sur des syllabes proéminentes ainsi que sur les syllabes adjacentes. L'accent est mis ici sur la configuration mélodique et sur les modalités d'alignement de ces mouvements intonatifs. Les phénomènes mélodiques plus globaux tels le registre ou le downstep ne sont pas encodés à ce niveau. La transcription des mouvements mélodiques se fait sur des bases perceptives et auditives, et non à partir d'une analyse acoustique de la fréquence fondamentale. Elle s'effectue à partir de l'écoute attentive d'une portion de signal correspondant au **domaine d'implémentation accentuel** (noté ID). L'extension de ce domaine varie selon les langues puisqu'elle dépend de la façon dont s'effectuent les associations tonales. En français, tout ID comprend i) la syllabe proéminente notée P ; ii) toutes les syllabes qui la précèdent jusqu'à la syllabe proéminente précédente ou jusqu'à la frontière d'un domaine intonatif majeur ; et iii) la syllabe qui la suit immédiatement. D'après cette définition, l'énoncé présenté dans la figure 1 se décompose en quatre ID, un pour chaque syllabe proéminente marquée par un mouvement mélodique.

Figure 2: Exemple de segmentation en IDs

Comme le montre la figure 2, deux ID consécutifs partagent généralement une syllabe. Ainsi la syllabe *de*



dans *le village de Beaulieu* appartient aux deux premiers IDs. Les IDs situés en début et en fin de domaines intonatifs ont respectivement une syllabe initiale et une syllabe finale qui n'apparaît dans aucun

autre ID. Les niveaux retenus pour l'encodage mélodique sont *haut* (H ou h), *moyen* (M ou m) et *bas* (L ou l) et sont tous relatifs. En outre, ces niveaux sont notés en lettres majuscules dès lors que leur cible est alignée sur le noyau d'une syllabe proéminente. La séquence *lHm*, par exemple, indique un mouvement mélodique montant du niveau bas vers le niveau haut, ce dernier étant atteint sur la syllabe marquée P. Ce mouvement est ensuite suivi d'une légère descente. Notons que les étiquettes représentent des valeurs relatives, puisqu'elles dépendent des réalisations au sein d'un unique ID.

Les phénomènes mélodiques se réalisant sur des empans qui dépassent le cadre d'un ID sont encodés sur la tire *Phénomènes phonétiques globaux*. Ainsi, les resettings qui sont parfois réalisés lorsque le locuteur introduit un nouveau thème de discours seront notés R sur cette tire. De même, un phénomène de downstep se réalisant au sein d'un domaine intonatif sera noté D.

Lorsque les événements notés sur les tires *Phénomènes phonétiques locaux* et *Phénomènes phonétiques globaux* sont considérés comme phonologiques, ils sont aussi encodés sur la tire *Phonologie*. La façon d'encoder les phénomènes mélodiques à ce niveau d'annotation dépend de l'inventaire phonologique des mouvements mélodiques contrastifs retenu pour la langue (ou variété) à transcrire. IVTS propose simplement un ensemble de primitives tonales parmi lesquels chaque transcripteur peut faire son choix (cf. [8] pour une discussion plus détaillée).

(2) Ensemble d'étiquettes pour le niveau phonologique

Tons		*Modifieurs*	*Frontières*	
H*	H	^ : upstep	%H	H%
L*	L	! : downstep	%L	L%
+		>: propagation	%	%

Pour finir, la tire *commentaires* est utilisée pour noter les remarques diverses.

2.2 Pourquoi quatre tires pour l'encodage prosodique ?

Recourir à quatre tires distinctes pour annoter différents aspects de la réalisation prosodique offre de nombreux avantages. Tout d'abord, en utilisant explicitement quatre niveaux pour annoter les phénomènes prosodiques, le système permet de décrire les variations dialectales de nature taxonimique. Comme les différences entre les langues peuvent affecter la distribution des syllabes proéminentes, la réalisation phonétique des contours intonatifs ou l'inventaire phonologique des primitives tonales, elles peuvent être encodées directement sur la tire dont elles relèvent.

Deuxièmement, le recours à plusieurs niveaux d'annotation prosodique permet d'obtenir plus d'uniformité et de transparence dans la réalisation des transcriptions. Analyser le système intonatif d'une langue impose de faire des choix sur i) le caractère contrastif ou non d'un mouvement mélodique ; ii) l'économie et la pertinence du système dans son ensemble, etc. Si la langue étudiée n'a pas encore été décrite, le recours aux différentes tires permet d'élaborer des hypothèses sur lesquelles il sera toujours possible de revenir par la suite.

Troisièmement, de par leur uniformité, les systèmes de transcription prosodiques dérivés de IViE permettent de mener facilement des comparaisons entre des langues, des dialectes (ou variétés) d'une même langue, voire entre des styles. Un ensemble d'étiquettes et de primitives est en effet proposé pour chaque tire, le transcripteur pouvant alors les combiner pour décrire telle ou telle variété, voire même en ajouter d'autres. Il est clair que les choix retenus au niveau phonologique résulteront souvent d'hypothèses théoriques particulières. Néanmoins, l'architecture globale du système et l'accès aux autres tires doivent faciliter les comparaisons.

Quatrièmement, la distinction entre les phénomènes phonétiques locaux et les phénomènes phonétiques globaux permet i) d'encoder les phénomènes discursifs ; ii) de traiter les langues dont le système intonatif ne repose pas une association entre événement tonal et proéminence syllabique.

2.3 Taxonimie des différences prosodiques

2.3.1 Différences cross-dialectales

L'étude pilote que nous avons menée sur trois variétés de français a confirmé le fait qu'un système dérivé de IViE est capable de révéler un grand nombre de différences taxonimiques entre dialectes, et cela dans des langues aussi différentes que le français ou l'anglais (cf. [8] et [9]).

Bien que limitée, notre étude [8] a montré qu'au niveau rythmique, les variétés de français parlées en Alsace, et probablement aussi en Belgique, se distinguent de nombreuses autres variétés de français par le fait que les syllabes pénultièmes en position finale de domaine intonatif majeur sont souvent proéminentes, et cela même si les syllabes finales le sont aussi (cf. l'étiquette P associée à la syllabe [e] de *émoi* dans la fig. 1). En français standard, les syllabes finales métriquement distinguées ne sont que très rarement précédées de syllabes proéminentes, et si elles le sont, cela se caractérise généralement par un mouvement mélodique (contra l'exemple de la fig. 1). Les parlers méridionaux, où les syllabes pénultièmes peuvent aussi être proéminentes, diffèrent à la fois du français standard et des variétés d'Alsace dans la mesure où i) les proéminences sur les syllabes pénultièmes n'apparaissent que lorsque les mots se terminent par un schwa comme dans *village* (non prononcé dans les variétés non méridionales) ; et ii) l'apparition de telles proéminences n'est pas conditionnée par la présence d'une frontière de domaine prosodique majeur, comme cela semble être le cas en Alsace. Ainsi, si nous comparons les productions du mot *village* dans les variétés de Marseille et de Liège, nous voyons que, dans les deux cas, la syllabe proéminente a pour noyau le [a]. Mais pour le locuteur de Liège, cette syllabe est

également la dernière syllabe du mot, tandis que pour celui de Marseille, ce n'est pas le cas. Ces différences sont clairement encodées dans IVTS au niveau de la tire *phénomènes phonétiques locaux*.

Au niveau phonologique, les dialectes du français sont également très différents les uns des autres. La variété parlée en Belgique montre clairement un contraste entre des mouvements mélodiques montant et descendant à l'intérieur d'un domaine intonatif (pitch accent H* vs. L*). En français standard, les mouvements descendant de ce type apparaissent plutôt en position finale de domaine intonatif (cf. [10 et [11]).

2.3.2 Différences entre langues

Les résultats de nos études sur les variations dialectales du français montrent que les différences relèvent de plusieurs niveaux d'analyse. Cela était également le cas pour l'Anglais britannique (cf. [9]). Ainsi, une même catégorie phonologique peut être réalisée différemment au niveau phonétique. Les recherches sur l'Anglais britannique ont par exemple montré qu'un mouvement mélodique descendant sur un mot comme *shift* est tronqué dans les variétés de Leeds, et compressé dans celles de Cambridge. Dans IViE, ces différences se reflétaient dans des différences d'étiquetage sur la tire *phénomènes phonétiques locaux* : lH-m était retenu dans le premier cas, et mH-l dans le second.

Les étiquettes retenues pour IViE ne sont pas les mêmes que celles utilisées pour le français (cf. [8]). Cela s'explique par le fait que les domaines pertinents pour rendre compte de la prosodie de chacune des deux langues sont différents. De la même façon, l'inventaire des formes phonologiques retenues pour encoder la phonologie de chacune des langues diffère. Malgré tout, les systèmes de transcription utilisés pour traiter ces deux langues reposent sur les mêmes principes, ce qui favorise donc les études comparatives.

Les comparaisons entre langues sont également intéressantes à mener dès lors que l'on s'intéresse à la question des transferts dans l'acquisition/ apprentissage des langues secondes. Considérons le cas d'apprenants espagnols ayant atteint un niveau intermédiaire en anglais britannique. Il est fréquent qu'ils produisent des mouvements mélodiques descendant en position prénucléaire dans un contexte approprié, mais avec un alignement incorrect qui relève d'une erreur d'implémentation phonétique : l'alignement est en effet celui attendu pour les mouvements montants en anglais britannique standard (données extraites d'une étude menée par Dolores Ramírez-Verdugo, UAM, Madrid et B. Post).

3. Conclusion et discussion

Le système de transcription IVTS permet d'encoder de façon discrète les événements prosodiques sur plusieurs niveaux, qu'ils soient locaux ou globaux. Cette façon de procéder pour analyser les phénomènes prosodiques offre de nombreux avantages comme la transparence, la flexibilité et la normalisation : autant d'éléments essentiels pour mener à bien des études comparatives. Dans ce papier, nous avons essayer de montrer comment un système comme IVTS permet i) de comparer des variétés du français et de l'anglais ; ii) d'identifier des difficultés d'acquisition de l'anglais langue seconde chez des hispanisants.

Nous sommes conscients que le prototype présenté ici nécessite d'être testé sur des données plus variées, et cela afin de vérifier si l'architecture globale du système est satisfaisante ou nécessite d'être modifiée. C'est ce que nous allons faire prochainement en travaillant sur d'autres variétés du Français et sur des variétés de l'Occitan. Par ailleurs, nous envisageons d'étudier l'apport que pourrait constituer un précodage automatique des données avec des logiciels comme Momel ([1]).

Bibliographie

[1] Campione, E.; Hirst, D.; Véronis, J., 2000. Automatic stylisation and symbolic coding of F0 implementations of the INTSINT model. In *Intonation. Research and Applications*, A. Botinis (ed.). Dordrecht : Kluwer.

[2] Beckman, M. E.; Hirschberg, J.; Shattuck-Hufnagel, S., 2005. The original ToBI system and the evolution of the ToBI framework. In *Prosodic Typology* : *The Phonology of Intonation and Phrasing*, S.-A. Jun (ed.). Oxford: Oxford University Press

[3] Grabe, E.; Post, B.; Nolan, F., 2001. Modelling intonational variation in English: The IViE system. In *Prosody 2000*, S. Puppel; G. Demenko (eds.). Poznan: Adam Mickiewicz University, 51-58..

[4] Pierrehumbert, J., 1980. The phonology and phonetics of English Intonation. Ph.D. thesis M.I.T.

[5] Ladd, D. R., 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[6] Auer, P.; Gilles, P., 2003. *Prosodic Variation in German between Areality and Pragmatics: an Overview*. Unpublished manuscript.

[7] Kochanski, G; Grabe E.; Coleman, J; Rosner, B., accepted. Loudness predicts prominence; fundamental frequency lends little. JASA.

[8] Post, B; Delais-Roussarie, E.; Simon, A-C., In press. Développer un système de transcription des phénomènes prosodiques. In *Bulletin PFC 4,* G. Caelen-Haumont ; A-C. Simon (eds.).

[9] Grabe, E.; Post, B., 2002. Intonational Variation in English. In *Proceedings of the Speech Prosody 2002 Conference*, B. Bel; I. Marlin (eds). Aix-en-Provence: Laboratoire Parole et Langage, 343-346.

[10] Post, B., 2000. *Tonal and phrasal structures in French intonation* (PhD thesis). The Hague: Holland Academic graphics.

[11] Martin, P., 1975. Analyse phonologique de la phrase française. *Linguistics* 146, 35-76.

Relations entre le bruit entachant les paramètres de contrôle de modèles non linéaires et le bruit mesuré à sa sortie

Michel Pitermann

Laboratoire Parole et Langage
Université de Provence, 29 av. R. Schuman - 13621 Aix-en-Provence Cedex 1, France
Tél. : +33 (0)4 42 95 36 26 - Fax : +33 (0)4 42 95 37 88
Émail : mpiter@lpl.univ-aix.fr

ABSTRACT

To carry out simulations by means of a nonlinear model, real data measured in our physical world are often used as values for the control parameters of the model. In this case, the output noise of the model should contain at least two components : (i) chaotic noise intrinsic to the model ; (ii) noise stemming from the measurements extrinsic to the model. A method to quantify the amplitude of chaotic noise was proposed in [2]. The present paper shows how the second noise component could be estimated for a biomechanical model of the face. The results show that despite its simplicity, the method correctly estimated the amplitude of the output noise of the model as a function of the noise present in the control parameters of the model.

1. INTRODUCTION

Les modèles non linéaires permettent de décrire de nombreux phénomènes, que ce soit en mécanique des fluides, en mécanique des solides, en acoustique, en perception, en sciences cognitives... De nombreuses simulations sont produites à l'aide de ces modèles. Malheureusement, il n'est pas toujours aisé de déterminer des paramètres de contrôle d'un modèle permettant d'atteindre les objectifs visés. Par exemples, en modélisation biomécanique de la langue ou du visage, quels muscles du modèle doit-on activer pour générer des mouvements correspondant à une phrase donnée ? On peut alors avoir recours à des mesures de grandeurs réelles. Par exemple, des mesures d'activité musculaire du visage ont été obtenues par électromyographie (*EMG*) pour un locuteur produisant un corpus de parole et des mouvements faciaux extrêmes. Ces mesures ont été utilisées comme valeurs d'activation musculaire d'un modèle biomécanique de visage afin de générer des animations d'aspect naturel [1]. Des activations musculaires potentielles peuvent aussi être obtenues à partir d'inversions d'enregistrements faciaux [4, 3]. Dans tous ces cas, les valeurs d'EMG étaient bruitées. Deux types de bruits étaient alors attendus à la sortie du modèle : (i) un bruit d'origine chaotique intrinsèque au modèle provenant des caractéristiques non linéaires du modèle ; (ii) un bruit provenant du bruit présent dans les paramètres de contrôle du modèle, donc extrinsèque au modèle.

Une méthode pour quantifier l'amplitude du bruit chaotique dans les modèles non linéaires a été proposée dans [2]. Bien que la méthode ait été présentée dans le cadre d'un modèle biomécanique de visage, elle peut être utilisée pour détecter le chaos dans n'importe quel modèle non linéaire. Cette méthode est simple, efficace et peu coûteuse à mettre en œuvre. Malheureusement, elle ne donne aucune indication sur le bruit issu de causes extrinsèques au modèle. Il s'agit d'une limitation importante en pratique car la méthode ne permet pas de prédire l'amplitude du bruit total attendu à la sortie d'un modèle, même lorsque l'on connaît le niveau de bruit à son entrée. Par exemple, en animation faciale, le bruit total mesuré à la sortie d'un modèle de visage peut être visible sous la forme d'une vibration de la peau. Il est donc important de pouvoir estimer le bruit total produit à la sortie du modèle. Par exemple, si l'on évaluait le seuil en dessous duquel une vibration de peau était imperceptible, la connaissance de la relation liant le bruit présent à l'entrée du modèle et le bruit mesuré à sa sortie permettrait de déterminer le niveau de bruit maximal admissible à l'entrée du modèle pour que la vibration de peau reste invisible.

L'article présenté ici complète [2] en proposant une méthode pour tenter de déterminer la relation quantitative liant le bruit présent dans les paramètres de contrôle du modèle au bruit mesuré à sa sortie. A nouveau, plutôt que d'offrir une présentation générale abstraite de la méthode, elle sera détaillée par son application sur le même modèle biomécanique de visage que celui qui a été utilisé dans [2]. Ensuite, la détermination du bruit maximal autorisé dans les paramètres de contrôle pour maintenir le bruit de sortie (la vibration de peau) en dessous d'un seuil choisi sera discutée. Bien que la méthode ne puisse donner satisfaction pour tous les modèles non linéaires, sa simplicité et le faible coût de sa mise en œuvre en fait une bonne candidate pour un premier essai.

2. MÉTHODE

L'idée générale de la méthode est d'émettre une hypothèse sur la forme de l'interaction entre le bruit intrinsèque et le bruit extrinsèque au modèle, puis d'émettre une autre hypothèse sur la manière dont le bruit d'entrée est propagé jusqu'à la sortie du modèle. Ensuite, on détermine le niveau de bruit mesuré à la sortie du modèle pour des paramètres de contrôle exempts de bruit. Enfin, on réalise les mêmes mesures de bruit à la sortie du modèle pour les mêmes valeurs des paramètres de contrôle, mais après avoir entaché ces paramètres par du bruit de différentes amplitudes. On peut alors vérifier les hypothèses de départ grâce à l'analyse de la relation entre le bruit total mesuré à la sortie du modèle et l'amplitude du bruit additionné aux paramètres de contrôle. Si les hypothèses de départ sont validées, les lois proposées liant le bruit à l'entrée et à la sortie du modèle sont correctes. Elles peuvent donc être utilisées telles quelles. Dans le cas contraire, il faut modifier les hypothèses de départ en fonction des résultats, et recommencer d'autres simulations afin de tenter de valider les nouvelles hypothèses. Passons à un exemple concret.

2.1. Le modèle de visage

Le modèle biomécanique de visage utilisé ici a été détaillé dans [1], seules ses principales caractéristiques seront résumées ici.

Le modèle était composé d'un module de mâchoire, d'un module de peau et d'un module de muscle. La mâchoire était décrite par une simple charnière contrôlée cinématiquement par un angle. La peau était composée d'un treillis de points massiques à trois couches aux propriétés mécaniques isotropes. Les 1434 masses du treillis étaient connectées par près de 6000 ressorts amortis non linéaires. Les muscles étaient modélisés par une formulation standard de type Hill. Les activations de ces muscles et l'angle de la mâchoire constituaient les paramètres de contrôle du modèle.

Différents sous-modules du modèle de visage contribuaient à sa non-linéarité. Tout d'abord le modèle de muscle était non linéaire. Ensuite le modèle de peau contenait des ressorts non linéaires en compression, des ressorts linéaires par morceaux en extension (biphasiques), une contrainte de volume, la réaction des corps solides (crâne, dents, globes oculaires) et une force de restauration nodale (cf. Eq. (8) de [1] pour les détails).

Huit paires de muscles (un élément de chaque paire pour chaque profil du modèle de visage) étaient activés simultanément avec le même niveau d'énergie pour l'expérience décrite ici. Les huit paires étaient le levator labii superior, le levator anguli oris, le zygomatic major, le depressor anguli oris, le depressor labii inferior, le mentalis, l'orbicularis oris superior et l'orbicularis inferior.

Pour animer le modèle, une image était calculée tous les 1/60 s. Les équations non linéaires de mouvement des points massiques étaient résolues à l'aide de l'algorithme de Runge–Kutta standard en utilisant un pas constant de 50 itérations par image. L'algorithme de Runge–Kutta pouvant être instable numériquement, un détecteur d'instabilité numérique a été ajouté. Au moindre signe d'instabilité, le nombre de pas était doublé, puis la stabilité était réévaluée.

2.2. Hypothèses choisies et séparation des différentes composantes de bruit

La première hypothèse retenue fut que la composante de bruit mesurée à la sortie du modèle générée par le bruit présent dans les paramètres de contrôle (bruit d'entrée) était indépendante de toutes les autres sources de bruit (chaos, erreurs d'arrondis…) Par conséquent, la variance totale σ^2_{tot} du bruit de sortie était supposée être la somme de la variance σ^2_{ent} de bruit dû au bruit d'entrée et de la variance σ^2_{x} de bruit dû à toutes les autres sources de bruit :

$$\sigma^2_{\mathrm{tot}} = \sigma^2_{\mathrm{x}} + \sigma^2_{\mathrm{ent}} \qquad (1)$$

σ^2_{ent} représente donc une composante de variance de bruit de *sortie* du modèle. C'est la partie de variance de bruit de sortie dû au bruit présent à l'entrée du modèle, et non la variance du bruit présent à l'entrée.

La deuxième hypothèse était que le modèle de visage propageait le bruit d'entrée comme un système linéaire. Ainsi, si l'amplitude de ce bruit était multipliée par un facteur r, la composante de variance due à ce bruit serait multipliée par r^2 :

$$\sigma^2_{\mathrm{tot}_r} = \sigma^2_{\mathrm{x}} + r^2\sigma^2_{\mathrm{ent}} \qquad (2)$$

où $\sigma^2_{\mathrm{tot}_r}$ était la nouvelle variance totale de bruit lorsque le bruit d'entrée original était multiplié par un facteur r.

En soustrayant l'Eq. (1) de l'Eq. (2), on obtient :

$$\sigma^2_{\mathrm{ent}} = \frac{\sigma^2_{\mathrm{tot}_r} - \sigma^2_{\mathrm{tot}}}{r^2 - 1}. \qquad (3)$$

En utilisant 5 amplitudes de bruit d'entrée différentes et en utilisant la plus petite amplitude comme référence pour σ^2_{tot}, σ^2_{ent} pouvait être estimé 4 fois à l'aide de l'équation (3) avec $r = 2$, 3, 4 et 5. Si aucune différence significative n'était mise en évidence pour les 4 estimations de σ^2_{ent}, l'Eq. (2) pouvait être considérée comme valide puisqu'elle avait servi de base à l'Eq. (3). Une valeur moyenne de σ^2_{ent} pouvait alors être tirée de ces quatre estimations, puis l'Eq. (1) pouvait être utilisée pour estimer σ^2_{x}, puis l'Eq. (2) permettait de connaître l'impact d'un bruit d'entrée de n'importe quelle amplitude sur les résultats.

2.3. Analyses statistiques

Les trajectoires 3D de 11 points du modèle de visage ont été suivies au cours du temps afin d'analyser statistiquement la vibration de peau (cf. figure 1 pour les positions approximatives de ces 11 points).

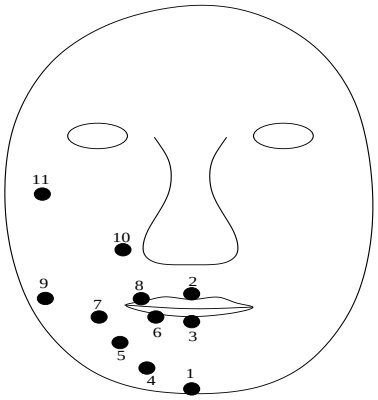


FIG. 1: Positions approximatives et numéros des points analysés (ellipses noires). Figure extraite de [2].

Quelques indices statistiques unidimensionnels ont été généralisés à trois dimensions afin de caractériser les trajectoires 3D des 11 points analysés. La position moyenne d'une trajectoire 3D v composée de n échantillons (x_i, y_i, z_i) était son centroïde μ_v :

$$\mu_v = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{n}\sum_{i=1}^{n} z_i \right) \qquad (4)$$

L'écart type σ_v d'une trajectoire 3D v était estimé par :

$$\sigma_v = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \|v_i\|^2 - \|\mu_v\|^2} \qquad (5)$$

où $\|\|^2$ était le carré de la norme d'un vecteur, c.à.d. la somme des carrés des coordonnées du vecteur. La différence entre deux séries chronologiques v et w était définie par :

$$\Delta_{vwi} = v_i - w_i \qquad \forall i \qquad (6)$$

où i représentait n'importe quelle coordonnée temporelle des séries chronologiques.

2.4. Stimuli et estimation du bruit

La constante 0.25 a été ajoutée à chaque élément de la série chronologique triangulaire (0, 1/12, 2/12, 3/12, 4/12, 5/12, 6/12, 5/12, 4/12, 3/12, 2/12, 1/12) pour obtenir des données appartenant à l'intervalle [0.25–0.75]. Cette série chronologique à été répétées 100 fois afin de créer une onde triangulaire de 1200 échantillons composée de la répétition de 100 triangles identiques.

L'onde triangulaire a été utilisée pour activer simultanément les 16 muscles sélectionnés du modèle. Il a été montré dans [2] que les trajectoires des 11 points analysés contenaient une composante systématique liée aux variations temporelles des activations musculaires et une composante de bruit chaotique. Pour déterminer l'amplitude du bruit chaotique, la sensibilité du modèle aux conditions initiales a été utilisée. La simulation a été répétée en déplaçant le point numéro 3 de 0.001 mm vers la droite pour la configuration initiale, les positions initiales des autres points étant identiques pour les deux simulations. Bien que quasi identiques au départ, les deux simulations se sont rapidement différenciées l'une de l'autre, la différence initiale augmentant exponentiellement jusqu'à varier « aléatoirement » autour d'une valeur type. L'amplitude du bruit chaotique a alors été estimée pour chacun des 11 points par $\sigma_{\Delta_{vw}}/\sqrt{2}$, où Δ_{vw} [Eq. (6)] représente la différence entre les deux trajectoires v et w du point pour les deux simulations (les détails sont présentés dans [2]). Ce bruit peut être considéré comme une estimation du σ_x des équations (1) et (2).

Ensuite, du bruit blanc a été ajouté à l'onde triangulaire pour 5 niveaux de bruit : 0.05, 0.10, 0.15, 0.20 et 0.25. Par exemple, une séquence de nombre pseudo-aléatoires de distribution uniforme comprise entre -0.1 et +0.1 a été utilisée comme bruit blanc d'amplitude 0.10. Toutes les valeurs d'activations musculaires étaient donc comprises entre 0 (pas d'excitation musculaire) et 1.0 (excitation musculaire maximale). Après activation des 16 groupes musculaires par une onde triangulaire bruitée, les trajectoires des 11 points analysés devaient contenir la même composante systématique de mouvement que pour les activations musculaires non bruitées plus une composante de bruit chaotique plus une composante de bruit résultant du bruit blanc d'entrée.

L'écart type σ_{tot_r} de bruit total de sortie du modèle a été estimé par $\sigma_{\Delta_{bn}}/\sqrt{2}$ pour chacun des 11 points analysés du modèle de peau et pour chaque niveau de bruit blanc ajouté aux contractions musculaires, où Δ_{bn} représente la différence entre la trajectoire b du point produite à l'aide des paramètres d'entrée bruités et la trajectoire n du même point produite à l'aide des paramètres d'entrée non bruités (la première simulation de l'expérience). Le bruit blanc de plus faible amplitude (0.05) a servi de référence pour σ^2_{ent} des équations (1) à (3) $\Rightarrow \sigma^2_{\mathrm{tot}}$ a été calculé par rapport à ce bruit blanc [Eq. (1)], et les 4 autres niveaux de bruit ont été utilisés pour les valeurs de $r = 2$, 3, 4 et 5.

3. RÉSULTATS

La figure 2 présente les estimations d'amplitude de bruit total à la sortie du modèle de visage lorsque les muscles ont été activés par une onde triangulaire entachée de bruit blanc pour différentes amplitudes de bruit. Pour passer de l'équation (1) à l'équation (3), il avait été pris pour hypothèse que la composante de variance du bruit total due au bruit blanc était indépendante de la composante de variance due aux autres sources de bruit. La figure 2 suggère que cette hypothèse pouvait être erronée. En effet, si l'hypothèse avait été correcte, le bruit total estimé à la sortie du modèle aurait toujours été plus important pour une onde triangulaire bruitée que pour une onde non bruitée. Par conséquent, la ligne pleine notée « 0.0 » de la figure 2 devrait toujours être sous les autres lignes à traits discontinus. Ce n'est pas le cas, surtout pour les points 2 et 6. Par conséquent, les équations (1) à (3) pourraient être inexactes et ne doivent être considérées que comme une première approximation.

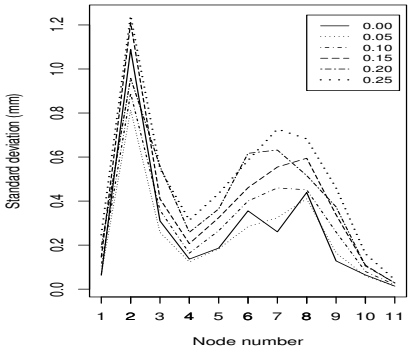


FIG. 2: Estimations de l'amplitude de bruit total à la surface de la peau du modèle de visage en fonction du numéro du point analysé (cf. Fig. 1 pour leur positionnement) lorsque les muscles du modèle ont été activés par une onde triangulaire entachée de bruit blanc pour différents niveaux de bruit. Les étiquettes de la légende allant de « 0.0 » à « 0.25 » indiquent l'amplitude du bruit blanc.

L'équation (3) a néanmoins été utilisée pour estimer quatre fois la composante de variance σ^2_{ent} due au bruit blanc d'amplitude 0.05 en utilisant $r = 2$, 3, 4 et 5 (les amplitudes des bruits blancs valant respectivement 0.10, 0.15, 0.20 et 0.25). Si les équations (1) à (3) étaient correctes, les 4 ensembles d'estimations de σ^2_{ent} devaient être équivalents. La figure 3 présente les 4 ensembles d'estimations d'amplitudes de bruit σ_{ent}. Une analyse de la variance à deux facteurs (« numéro du point » et r) a été réalisée pour σ^2_{ent} afin de tester son indépendance par rapport à r. Les estimations de σ^2_{ent} dépendaient bien du « numéro du point » [$F(10, 30) = 6.51; p < 1e-4$], mais pas de r [$F(3, 30) = 1.84; p = 0.16$] au niveau de risque 0.05. Cela signifie que les 4 ensembles d'estimations de σ^2_{ent} ne différaient pas significativement au niveau 0.05. Par conséquent, les équations (1) et (2) modélisaient correctement, en première approximation, la contribution du bruit blanc au bruit total estimé à la sortie du modèle. Nous pouvons donc considérer que la composante de variance de vibration de peau attribuable à un bruit blanc présent dans les activations musculaires variait linéairement avec le carré de l'amplitude du bruit blanc.

4. DISCUSSION

Les données de la figure 2 et 3 associées à la loi des carrés (2) permettent d'estimer l'impact sur la vibration de peau d'un bruit blanc de n'importe quelle amplitude entachant des données EMG. Par exemple, d'après la figure 2, l'amplitude de bruit total à la surface de la peau en absence de bruit blanc valait à peu près 0.3 mm pour les points 6, 7 et 8 (cf. ligne pleine étiquetée « 0.00 » pour ces points). Il s'agit d'une estimation du σ_x des équations

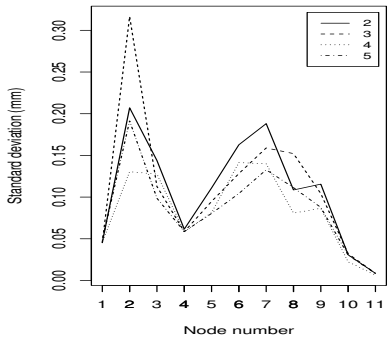


FIG. 3: Estimations de l'amplitude de la composante σ_{ent} de bruit de sortie due au bruit blanc d'amplitude 0.05 présent dans l'EMG en fonction du numéro du point analysé (cf. Fig. 1 pour leur positionnement). Les étiquettes « 2 » à « 5 » de la légende indiquent les valeurs de r utilisées dans l'équation (3) pour produire les estimations.

(1) et (2) $\Rightarrow$ la variance σ_{x}^2 valait à peu près 0.09 mm^2. La figure 3 indique qu'un bruit blanc d'amplitude 0.05 (5 % de la contraction maximale d'un muscle) dans les données EMG entraîne une contribution à la vibration de peau de l'ordre du dixième de mm, donc une contribution à la variance de bruit total de l'ordre de 0.01 mm^2. Si des nouvelles mesures EMG étaient entachées d'un bruit blanc d'amplitude de l'ordre de 0.25 (25 % de l'activation maximale d'un muscle), alors ce bruit aurait 5 fois l'amplitude du bruit blanc de référence (0.05) $\Rightarrow r = 5$ dans l'équation (2). Ce nouveau bruit blanc apportera donc une contribution de variance de bruit 25 fois supérieure à celle du bruit original (loi en r^2) $\Rightarrow$ la variance totale de vibration de peau $\sigma^2_{\mathrm{tot}_5} = \sigma^2_{\mathrm{x}} + 25\,\sigma^2_{0.01}$ [Eq. (2)] $\Rightarrow \sigma^2_{\mathrm{tot}_5} = 0.09 + 0.25 = 0.34\,\mathrm{mm}^2 \Rightarrow \sigma_{\mathrm{tot}_5} = \sqrt{0.34} \approx 0.6\,\mathrm{mm}$, ce qui peut être vérifié pour les points 6, 7 et 8 de la ligne pointillée étiquetée 0.25 de la figure 2. L'opération peut être recommencée pour n'importe quel point et pour n'importe quel niveau de bruit blanc présent dans des données EMG.

La quantité de bruit acceptable dans des mesures EMG avant qu'une vibration de peau ne dépasse un certain seuil est calculable à l'aide du même raisonnement. Imaginons par exemple que sous les conditions de lumière des animations, une vibration de peau devienne visible à partir de 0.4 mm d'amplitude. On souhaiterait donc que $\sigma^2_{\mathrm{tot}_r} < 0.16\,\mathrm{mm}^2 \Rightarrow \sigma^2_{\mathrm{tot}_r} = 0.09 + r^2\,0.01 < 0.16$ [Eq. (2)] $\Rightarrow r < \sqrt{7} \approx 2.65 \Rightarrow$ l'amplitude de bruit blanc acceptable dans les mesures EMG ne peut excéder $\sqrt{7}\,\sigma_{\mathrm{ent}} = \sqrt{7}\,0.05 \approx 0.13$ (13 % de la contraction maximale d'un muscle). La méthode permet donc de déterminer le niveau de « propreté » nécessaire des mesures physiques si l'on souhaite maintenir le niveau de bruit total à la sortie du modèle sous un seuil choisi à priori.

Un inconvénient de la méthode est de devoir expliciter l'interaction entre les variances des différentes composantes de bruit [Eq. (1)] ainsi que la loi de propagation du bruit d'entrée dans le modèle [Eq. (2)]. Heureusement, les hypothèses choisies ont été satisfaisantes en première approximation pour le modèle de visage analysé ici. En effet, bien que l'hypothèse d'indépendance a été légèrement mise en défaut (cf. commentaire dans le texte relatif à la figure 2), les lois fonctionnaient correctement en première approximation (cf. résultats de l'analyse de la variance et la vérification de l'amplitude du bruit de sortie pour un bruit blanc d'amplitude 0.25 entachant les activations musculaires). Lorsque des hypothèses aussi simples ne sont pas du tout satisfaisantes, le graphique de la relation entre l'amplitude du bruit entachant les paramètres de contrôle et l'amplitude du bruit à la sortie du modèle peut être produit. On peut alors tenter de tabuler ou modéliser cette relation puis appliquer les mêmes raisonnements qu'ici, même si le succès n'est pas garanti.

Il faut aussi noter que du bruit coloré peut entacher les paramètres de contrôle du modèle et ne pas suivre exactement la loi de propagation du bruit blanc. Il faut alors réaliser les simulations avec un bruit de même couleur que celui attendu dans les paramètres de contrôle.

Malgré les limitations de la méthode, les résultats obtenus pour le modèle biomécanique de visage sont encourageants. On aurait pu craindre que les hypothèses choisies pour les équations (1) et (2) soient trop simples en regard de la non-linéarité du modèle. En effet, il a été montré dans [2] que les non-linéarités du modèle sont suffisantes pour entraîner un bruit chaotique dont l'amplitude vaut à peu près 10 % de l'amplitude d'un mouvement de visage produit par une contraction maximale des 16 groupes musculaires. Malgré l'importance relative de ces non-linéarités, l'analyse de la variance présentée ici montre que les hypothèses retenues étaient satisfaisantes pour produire des formules constituant de bonnes approximations des relations liant les bruits intrinsèques et extrinsèques au modèle au bruit mesuré à sa sortie.

5. CONCLUSION

Les modèles non linéaires produisent généralement du bruit chaotique intrinsèque aux modèles. Une méthode générale destinée à estimer l'impact de ce bruit sur les simulations produites a été présentée dans [2]. Le présent article complète cette méthode en proposant une technique destinée à estimer l'impact du bruit extrinsèque au modèle sur les simulations. Il faut d'abord émettre des hypothèses sur l'interaction entre les différents types de bruits ainsi que sur leur propagation dans le modèle. Ensuite, quelques simulations simples et peu coûteuses à mettre en œuvre suffisent parfois à valider les lois retenues. Grâce à ces lois, il peut être possible de déterminer le niveau de bruit maximal acceptable dans les paramètres de contrôle du modèle pour que le bruit total généré à sa sortie soit inférieur à un seuil choisi.

RÉFÉRENCES

[1] Jorge C. Lucero and Kevin G. Munhall. A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America*, 106(5) :2834–2842, 1999.

[2] Michel Pitermann. Chaos dans la modélisation des tissus mous. In *XXVe Journées d'Étude sur la Parole*, pages 401–404, Fez, Maroc, 2004.

[3] Michel Pitermann and Kevin G. Munhall. An inverse dynamics approach to face animation. *The Journal of the Acoustical Society of America*, 110(3) :1570–1580, 2001.

[4] Demetri Terzopoulos and Keith Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6) :569–579, 1993.

Modélisation physique des cordes vocales : Comment tester la validité des modèles ?

Nicolas Ruty, Annemie Van Hirtum, Xavier Pelorson

Institut de la Communication Parlée
46, Avenue Félix Viallet, 38 031 Grenoble Cedex 1, France
ruty@icp.inpg.fr
http://www.icp.inpg.fr

ABSTRACT

An experimental set-up and human vocal folds replica able to produce self sustained oscillations is presented. The aim of the set-up is to assess the relevance and the accuracy of theoretical vocal folds models. The applied reduced mechanical models are a variation of the classical two-mass model. The airflow is described as a laminar flow with flow separation. The influence of a downstream resonator is taken into account. The oscillation pressure threshold and fundamental frequency are predicted by applying a linear stability analysis to the mechanical models. The measured frequency response of the mechanical replica together with the initial (rest) area allows to determine the model parameters (spring stiffness, damping, geometry, masses). Validation of theoretical model predictions to experimental data shows the relevance of low order models in gaining a qualitative understanding of phonation. However quantitative discrepancies remain large due to an inaccurate estimation of the model parameters and the crudeness in either flow or mechanical model description. As an illustration it is shown that significant improvements can be made by accounting for viscous flow effects.

1. INTRODUCTION

La phonation peut être vu comme le résultat d'une interaction complexe entre l'écoulement d'air provenant des poumons et les tissus déformables que constituent les cordes vocales. L'auto-oscillation de celles-ci est la principale source pour la production des sons voisés.

La modélisation physique de ce phénomène a des applications telles que la synthèse de parole et l'étude de pathologies vocales. Les modèles numériques complets, s'appuyant des modèles à éléments finis des tissus (Hunter & col. [1]), sont peu utilisés, pour plusieurs raisons. La complexité numérique de cette approche impose des temps de calculs importants et des phénomènes tels que la collision des cordes vocales, des conditions limites instationnaires, ou la présence de turbulence dans l'écoulement d'air sont difficiles à prendre en compte.

L'utilisation de modèles distribués est fréquente dans le domaine de la synthèse de parole (Ishizaka & Flanagan[2]), mais aussi pour l'étude des pathologies vocales (Lous & col.[3]). Peut-on alors les utiliser pour rendre compte des interactions qui interviennent lors de la phonation, en incluant les aspects aérodynamiques, biomécaniques et acoustiques?

Concernant l'écoulement d'air, des travaux de recherche ont été menés et ont permis de tester des théories complexes. On distingue les études « *in-vivo* », et les études « *in-vitro* », effectuées sur des maquettes plus ou moins proches de la réalité physiologique. Ainsi, des mesures d'écoulements sur des maquettes rigides de cordes vocales ont été réalisées par van den Berg & col.[4]. Récemment, des écoulements plus complexes et réalistes ont été obtenus sur des maquettes de cordes vocales oscillantes(Titze & col.[5]). Plus particulièrement, les maquettes capables de produire des oscillations auto-entretenues présentent l'intérêt de pouvoir tester la description de l'écoulement mais aussi l'interaction avec des structures mécaniques et un résonateur acoustique.

Mais fondamentalement, quels liens ont les modèles théoriques distribués avec la physiologie des cordes vocales ? Si certains paramètres des modèles théoriques, comme la pression sous glottique ou la géométrie de la glotte, peuvent être directement reliés aux observations faites sur l'être humains, d'autres n'ont pas de lien direct avec la réalité. L'existence d'une relation entre ces paramètres et la physiologie de l'appareil phonatoire est donc un réel problème. D'important travaux (Svec & col.[6]) de mesures *in-vivo* des propriétés mécaniques des cordes vocales humaines ont été menés. Les résultats obtenus sont utiles pour déterminer les paramètres mécaniques des modèles théoriques, mais ne permettent pas de tester la validité des modèles.

Nous présentons un dispositif expérimental permettant le test des modèles théoriques de cordes vocales. En effet, si les modèles théoriques ne peuvent pas simuler le comportement de maquettes plus simples que la réalité, comment justifier qu'ils puissent simuler le comportement de vraies cordes vocales. Ce dispositif consiste en une maquette de cordes vocales capable d'auto-osciller. L'avantage des maquettes est qu'elles sont contrôlables et que les mesures sont répétables et reproductibles. Un système optique permet de mesurer la réponse mécanique de la maquette, mais aussi son déplacement durant les oscillations. On utilise ce dispositif pour tester le modèle théorique à deux masses d'Ishizaka & Flanagan[2].

2. Modélisation théorique:

2.1 Modèle d'écoulement

La géométrie de la glotte est décrite sur la figure 1. Le modèle d'écoulement s'appuie sur l'hypothèse d'un écoulement laminaire, incompressible et quasi-stationnaire à travers la glotte. La pression sous glottique P_{sub} est supposée constante. Lorsque la glotte forme un canal divergent, le point de séparation de l'écoulement a une position déterminée par un critère *ad-hoc* As(t) = 1.2*min(A(x,t)) avec A(x,t) l'aire de la section le long de la glotte, et As(t) l'aire de la glotte au point de séparation. Le débit U_g est supposé constant. La distribution de pression P(x,t) le long de la glotte s'écrit

$$P(x,t) = P_{sub} - (P_{sub} - P_{sup\,ra})\left(\frac{A_s^2}{A^2(x,t)}\right), \text{ si } x < x_s \quad (1)$$

$P(x,t) = P_{\sup ra}$, si $x > x_s$

où ρ est la densité de l'air, U_g le débit volumique, $P_{\sup ra}$ la pression supra-glottique.

2.2 Modèle biomécanique

Les propriétés mécaniques des cordes vocales sont décrites par le modèle théorique représenté sur la figure 1. La largeur de la glotte est supposée égale à Lg. Nous supposons que les deux cordes vocales ont un mouvement symétrique. En appliquant le principe de la résultante dynamique aux deux masses, on obtient :

$$\begin{aligned} \frac{m}{2}\frac{\partial^2 H_1(t)}{\partial t^2} &= -k(H_1(t)-H_{10}) - k_c(H_1(t)-H_{10}-H_2(t)+H_{20}) - r\frac{\partial H_1(t)}{\partial t} + F_1(P_{sub}, P_{supra}, H_1, H_2) \\ \frac{m}{2}\frac{\partial^2 H_2(t)}{\partial t} &= -k(H_1(t)-H_{20}) - k_c(H_2(t)-H_{20}-H_1(t)+H_{10}) - r\frac{\partial H_2(t)}{\partial t} + F_2(P_{sub}, P_{supra}, H_1, H_2) \end{aligned} \quad (2)$$

où m est la masse effective d'une corde vocale, , k et k_c sont les constantes de raideurs des ressorts, r est l'amortissement, d la longueur de la glotte, $H_1(t)$ et $H_2(t)$ indiquent l'ouverture au niveau des deux masses, H_{10} et H_{20} les positions de repos, F_1 et F_2 les forces de pression suivant l'axe des ordonnées.

Le phénomène de contact entre les cordes vocales est pris en compte, comme indiqué par Ishizaka & Flanagan[2] . La source glottique ainsi modélisée constitue, par les fluctuations de débit qu'elle engendre, une source de pression acoustique. Du fait des dimensions de la glotte par rapport au conduit vocal, cette source est considérée comme ponctuelle, ce qui justifie la coexistence avec le modèle de résonateur acoustique couplée à cette source.

2.3 Modèle acoustique

L'acoustique du résonateur aval est modélisée par l'équation :

$$\frac{\partial^2 \psi(t)}{\partial t^2} + \frac{\omega_A}{Q_A}\frac{\partial \psi(t)}{\partial t} + \omega_A^2 \psi(t) = \frac{Z_A \omega_A}{Q_A} u \quad (3)$$

où $\frac{\partial \psi(t)}{\partial t} = p$, p est la pression acoustique à l'entrée du conduit vocal, ω_A la pulsation de résonance du conduit vocal, Q_A le facteur de qualité de cette résonance, Z_A est l'impédance à la pulsation de résonance ω_A, u est le débit acoustique.

3. Dispositif expérimental:

3.1 Description

Une maquette de cordes vocales est fixée sur un réservoir de 0.75m³ (voir figure 2), alimenté par un compresseur. La pression en amont de la maquette peut varier de 0 à 3000Pa. La maquette de cordes vocales (voir figures 2 et 3) consiste en deux demi cylindres métalliques (diamètre 12.5mm) recouverts de tubes en latex (diamètre 11mm, +/- 0.1mm, épaisseur 0.2mm, +/-10%). Elle est remplie d'eau sous pression (pression interne Pc). Cette pression a une influence directe sur les caractéristiques mécaniques et géométrique de la maquette. Pour de faibles valeurs de Pc (~2500Pa), l'ouverture de la maquette est grande, et le latex est peu tendu, alors que pour de fortes valeurs de Pc (~6000Pa), les deux tubes sont en contact et fortement tendus. Un résonateur acoustique est connecté en aval de la maquette de cordes vocales. Deux résonateurs différents, de section circulaire (diamètre 25mm, longueur 250mm et 500mm) ont été utilisés.

Des mesures de pression sont réalisées avec deux capteurs de pression Kulite XCS-0.93-0.35-Bar-G. Le premier est placé en amont de la réplique de cordes vocales, et le second à l'extrémité du résonateur acoustique. Des mesures de déplacements sont réalisées au moyen d'un système optique composé d'une diode laser, et d'un ensemble de lentilles. Le faisceau traverse le réservoir d'air, il est ensuite modifié par les mouvements la maquette des cordes vocales. La variation d'intensité lumineuse est mesurée grâce une photo diode (précision de 0.01mm).

3.2 Détermination des paramètres du modèle théorique

Si certains paramètres peuvent être connus directement (Lg et d), d'autres doivent être estimées.

Ainsi, la masse m, utilisée dans la description du modèle biomécanique théorique, est estimée grâce à la quantité d'eau présente à l'intérieur du tube en latex de la maquette. Ainsi on a :

$$m_{cv} = \rho_e L \frac{\pi d_l^{\,2}}{8} \quad (4)$$

où ρ_e est la masse volumique de l'eau, L la largeur de la maquette, d_l le diamètre d'un tube (1.1mm). Ainsi, on a $m_{cv} = 2.29\,g$.

Les raideurs des ressorts et les amortissements sont estimés par la mesure de réponses mécaniques effectuées sur la maquette. On crée une excitation acoustique (voir figure 2). La réponse de la maquette est alors tracée en fonction de la fréquence d'excitation, variant entre 100Hz et 400Hz par pas de 1Hz. Ce protocole est répété pour des pressions internes variant de 500Pa à 6500Pa par pas de 500Pa.

Un exemple typique de réponse mécanique est présenté sur la figure 4. On peut noter que celle-ci est très similaire à celle de cordes vocales humaines (Svec & col[6]).

On peut extraire des réponses mécaniques les pulsations de résonances ω_0 et leurs facteurs de qualité associés Q_0. Ces paramètres sont reliés aux résonances naturelles du modèle théorique, donc aux constantes de raideurs et des constantes d'amortissements :

$$\omega_0 = \sqrt{\frac{2k}{m}}, \quad Q_0 = \frac{m\omega_0}{2r} \quad (5)$$

où k est la constante de raideur des ressorts, $m = m_{cv}/2$ est la masse effective d'une corde vocale, r est l'amortissement. La constante de raideur de couplage k_c est fixée à $k/2$.

4. Résultats et discussion:

4.1 Mesure des pressions de seuil d'oscillation

On impose une pression interne Pc dans la maquette de cordes vocales. L'ouverture initiale h_0 est mesurée en l'absence d'écoulement d'air. La pression en amont de la maquette est progressivement augmentée jusqu'à ce que des oscillations apparaissent. Cette première pression de seuil est notée **Pon-set**. On note la fréquence des oscillations. La pression d'alimentation de la maquette est diminuée jusqu'à ce que les oscillations cessent, pour une pression **Poff-set**. Cette opération est répétée pour une pression interne Pc variant de 500Pa à 6500Pa. Les résultats obtenus sont présentés sur les figures 5a et 5b.

On observe qu'en fonction de la longueur du résonateur aval, les oscillations apparaissent sur deux plages différentes de pression interne Pc. Ainsi, pour le résonateur court (25mm), les oscillations apparaissent pour de faibles pressions internes (Pc ϵ[500;2500]Pa), correspondant à de grandes ouvertures initiales. Pour le long résonateur (500mm), les oscillations apparaissent pour des pressions internes plus élevées (Pc ϵ[3500;6500]Pa). Dans ce cas, on observe que pour Pc=5000Pa, la pression de seuil d'oscillation passe par un minimum. Pour ce point particulier, l'ouverture initiale est presque nulle, i.e. les deux tubes en latex sont quasiment en contact. Ce minimum de pression, aussi observé par Titze et col.[5] peut être relié à la configuration optimale pour la production de sons voisés. Enfin, dans les deux configurations, un phénomène d'hystérésis est apparu, tel que décrit théoriquement par Lucero[7].

4.2 Analyse linéaire de stabilité des équations théoriques

Toutes les variables sont linéarisées. Sous l'hypothèse d'une pression sous-glottique constante et d'une pression supra-glottique nulle, les équations du modèle théorique peuvent s'écrire sous d'une représentation d'état avec $x = \left[h_1 \quad h_2 \quad \psi \quad \frac{\partial h_1}{\partial t} \quad \frac{\partial h_2}{\partial t} \quad \frac{\partial \psi}{\partial t} \right]$ pour vecteur d'état, on a alors $\frac{\partial x}{\partial t} = Mx$ où M est la matrice d'état :

$$M = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -\frac{2}{m}\left(k+k_c-2\frac{\partial F_1}{\partial H_1}\right) & \frac{2}{m}\left(k_c+2\frac{\partial F_1}{\partial H_2}\right) & 0 & -\frac{2r}{m} & 0 & \frac{4}{m}\frac{\partial F_1}{\partial P_{supra}} \\ \frac{2}{m}\left(k_c+2\frac{\partial F_2}{\partial H_1}\right) & -\frac{2}{m}\left(k+k_c-2\frac{\partial F_2}{\partial H_2}\right) & 0 & 0 & -\frac{2r}{m} & \frac{4}{m}\frac{\partial F_2}{\partial P_{supra}} \\ \frac{Z_A\omega_A}{Q_A}\frac{\partial U_g}{\partial H_1} & \frac{Z_A\omega_A}{Q_A}\frac{\partial U_g}{\partial H_2} & -\omega_A^2 & 0 & 0 & -\frac{\omega_A}{Q_A}+\frac{Z_A\omega_A}{Q_A}\frac{\partial U_g}{\partial P_{supra}} \end{bmatrix} \quad (6)$$

En étudiant les valeurs propres de M, on peut déterminer théoriquement la présence d'oscillations, ainsi que la fréquence de ces oscillations.

Dans la partie expérimentale, on a vu qu'imposer une pression interne Pc affectait non seulement la géométrie initiale, mais aussi les propriétés mécaniques. Pour une valeur donnée de Pc, la présence d'oscillations dépendait des variations de la pression en amont de la maquette. De même ici, pour un jeu de paramètres du modèle théorique donné (constante de raideur, amortissement, pression sous glottique, géométrie initiale), l'analyse linéaire de stabilité est réalisée pour des valeurs de pression sous-glottique P_{sub} variant de 0 à 1000Pa. Cette procédure est répétée pour chacun des jeux de paramètres (correspondant à une pression interne Pc) et pour chacun des deux résonateurs acoustiques, de longueurs respectives 250mm et 500mm. Les résultats sont tracés sur les figures 6a et 6b, sur lesquelles ils sont comparés aux données expérimentales.

4.3 Discussion

Les résultats prédits par le modèle théorique, montrent leur capacité à reproduire qualitativement ce qui est observé expérimentalement, mais avec une marge d'erreur importante.

Plus précisément, pour les fréquences fondamentales, la prédiction est assez précise (marge d'erreur inférieure à 10%). Les fréquences fondamentales augmentent suivant l'augmentation de pression interne Pc, ce qui est cohérent avec le fait qu'accroître cette pression a pour conséquence une augmentation de la tension de la maquette. En effet, accroître la tension signifie augmenter les constantes de raideur et donc faire croître les pulsations de résonance naturelle du modèle théorique.

Concernant les pressions de seuil, on observe que le modèle se comporte suivant le même ordre de grandeur. Globalement, la même forme pour les pressions de seuil en fonction de la pression interne Pc est obtenue, sauf dans le cas où l'ouverture initiale est nulle, i.e. lorsque il y a contact entre les deux cordes vocales. Dans ce cas, l'erreur de prédiction est conséquente. Quantitativement, la correspondance mesures/théorie est donc médiocre, sauf à proximité du seuil minimum.

Les écarts observés peuvent bien sur être dus aux modèles eux mêmes mais aussi à une mauvaise estimation de certains paramètres. Ainsi, la masse totale m_{cv} est estimée de façon géométrique. Cette estimation repose sur l'hypothèse d'une masse constante quelque soit la pression interne Pc. Or, pour tenir compte des variations liées à la quantité d'eau que contiennent les tubes en latex, une estimation de la masse, variable en fonction de l'ouverture initiale, doit être proposée :

$$m_{cv} = \rho_e L \frac{\pi\left(d_l + (h_{ref} - h_0)/4\right)^2}{8} \quad (7)$$

h_{ref} est une ouverture de référence valant 3mm, h_0 est l'ouverture initiale mesurée sur la maquette, variant en fonction de Pc.

Après cette correction, les pressions de seuil sont modifiées avec un ordre de grandeur de l'ordre de 10%. Ce paramètre de masse ne semble donc pas pouvoir expliquer seul les écarts observés précédemment. On teste alors l'influence d'un modèle d'écoulement visqueux. En ajoutant un terme de Poiseuille l'équation (1) devient :

$$\begin{aligned} P_{sub} - P(x,t) &= \frac{1}{2}\rho U_g^2\left(\frac{1}{A^2(x,t)}\right) - 12\mu L_g^2 U_g \int_{x_0}^{x} \frac{dx}{A^3(x,t)}, \text{ if } x < x_s \\ P(x,t) &= P_{supra}, \text{ if } x > x_s \end{aligned} \quad (8)$$

où μ est le coefficient de viscosité de l'air.

Cette modification (figures 6a et 6b) a un effet significatif lorsque l'ouverture initiale est nulle ou quasi-nulle. Le terme visqueux augmente significativement la correspondance mesures/théorie. Cependant, cela ne semble pas complètement suffisant pour expliquer les écarts entre mesures et prédictions théoriques.

5. Conclusion

Un dispositif expérimental permettant de tester des modèles théoriques de cordes vocales a donc été présenté. Cette approche est illustrée par le test d'un modèle à faible nombre de degrés de liberté, dont les paramètres sont reliés à des grandeurs mesurables et contrôlables expérimentalement. La comparaison des prédictions théoriques de ce modèle avec les données expérimentales a permis de conclure que malgré sa simplicité, le modèle théorique prédit qualitativement bien les données expérimentales, même si l'aspect quantitatif demande plus de développement.

Remerciements

Ces travaux ont été supportés en partie par le ministère de la recherche, par le biais d'une bourse de doctorant et dans le cadre du projet Franco-Allemand Popaart (CNRS-MAE). Nous souhaiterions aussi remercier Pierre Chardon pour la mise en place du dispositif expérimental, et Freek van Uittert pour ce qui concerne le système d'acquisition de données.

BIBLIOGRAPHIE

[1] E.J. Hunter, I.R. Titze, F. Alipour. A three-dimensional model of vocal fold abduction/adduction. *J. Acoust. Soc. Am.* 115:1747-57, 2004.

[2] K. Ishizaka, J.L Flanagan. Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords. *Bell Syst. Tech. Journal* 51:1233-1267, 1972.

[3] N.J.C Lous, G.C.J Hofmans, N.J Veldhuis, A. Hirschberg. A symetrical two-mass model vocal-fold model coupled to vocal tract and trachea, with application to prothesis design. *Acustica* 84:1135-1150, 1998.

[4] Jw. Van den Berg, J.T. Zantema, P. Doornenbal. On the air resistance and the Bernoulli effect of the human larynx. *J. Acoust. Soc. Am.* 29:625-631, 1957.

[5] I.R. Titze, S.S. Schmidt, M.R. Titze. Phonation Threshold pressure in a physical model of the vocal fold mucosa. *J. Acoust. Soc. Am.* 97:3080-3084, 1995.

[6] J.G. Svec, J. Horacek, F. Sram, J. Vesely. Resonance properties of the vocal folds: *In vivo* laryngoscopic investigation of the externally excited laryngeal vibrations. *J. Acoust. Soc. Am.* 108:1397-1407, 2000.

[7] J.C. Lucero. A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. *J. Acoust. Soc. Am.* 10:423-31, 1999.

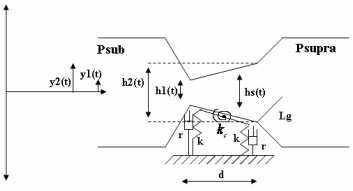


Figure 1 : Modèle à deux masses de cordes vocales, inclus dans une géométrie 2D pour le calcul de la distribution de pression.

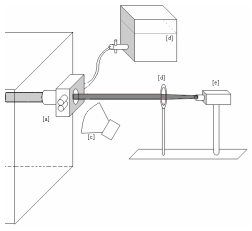


Figure 2 : Dispositif expérimental. [a] Maquette de cordes vocales. [b], Réservoir d'eau. [c], haut parleur. [d], lentille convergente. [e], photo diode.

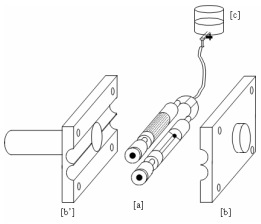


Figure 3 : Maquette de cordes vocales (Vilain et al., 2004). [a] demi cylindres en métal, recouverts de latex, puis remplis d'eau. [b], [b'], supports métalliques. [c] réservoir d'eau.

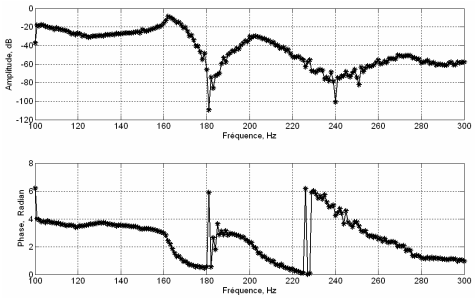


Figure 4 : Réponse mécanique de la maquette de cordes vocales, Pc=2000Pa. [a] amplitude en dB. [b] phase en radian.

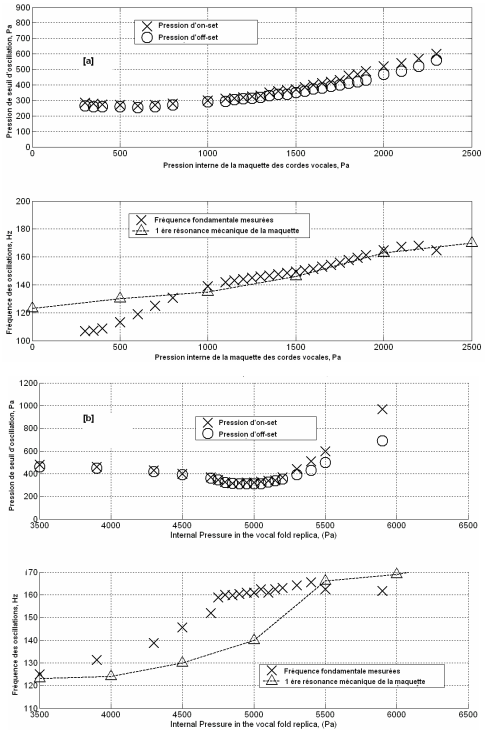


Figure 5 : Tracés expérimentaux des seuils de pression d'oscillation et des fréquences fondamentales d'oscillation. [a], pour un résonateur acoustique de 250mm de long. [b], pour un résonateur acoustique de 500mm de long.

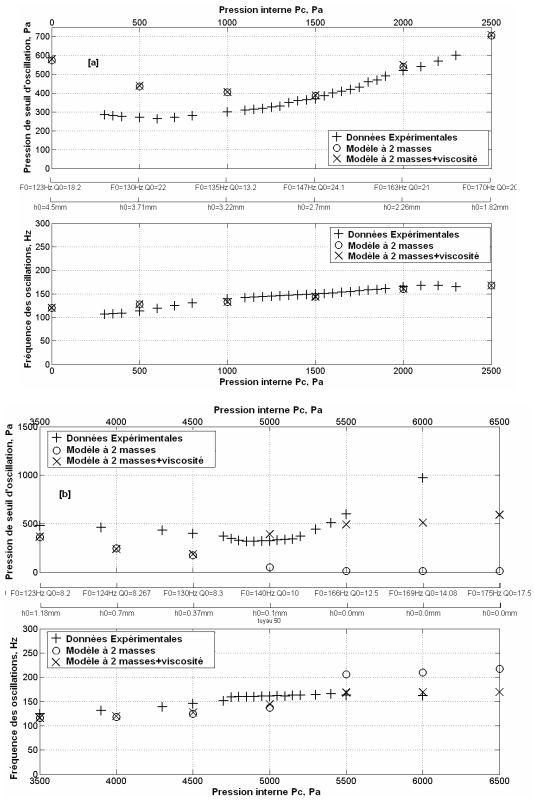


Figure 6 : Etude comparative entre les résultats de l'analyse de stabilité et les données expérimentales. [a], pour un résonateur acoustique de 250mm de long. [b], pour un résonateur acoustique de 500mm de long.

Analyse dynamique de la réduction vocalique en contexte CV à partir des pentes formantiques en arabe dialectal et en français.

Jalaleddin AL-TAMIMI

Laboratoire Dynamique du Langage (UMR 5596)
Institut des Sciences de l'Homme, 14 av. Berthelot – 69007 Lyon, France
Mél: Jalal-Eddin.Al-Tamimi@univ-lyon2.fr - http://www.ddl.ish-lyon.cnrs.fr/

ABSTRACT

Linear regression parameters (formant slopes and intercepts) are proposed to measure the degree of vowel reduction in 3 vowel systems: Moroccan Arabic, Jordanian Arabic and French. 10 speakers per language produced a list of vowels in C_1VC, C_1VCV, or C_1VC_2VC words, where C_1 or C_2 was /b/, /d/ or /k/. Our results show that the values of formant slopes and intercepts are dependent on: 1) the place of articulation of adjacent consonants, 2) vowel quality, and 3) the language's vowel system density. Discriminant analysis results show the possibility of language separation on the basis of *F*1, and *F*2 slopes, intercept values, and duration.

1. INTRODUCTION

En phonétique, on parle de réduction vocalique (RV) lorsque les valeurs cibles des voyelles (produites isolément) ne sont pas atteintes, selon le contexte. Les voyelles ont tendance à être "réduites" vers une forme *plus ou moins* centralisée (sur les deux axes *F*1 et *F*2) se rapprochant ainsi d'un schwa [ə] (Lindblom [1]). À débit rapide et en position non-accentuée, les voyelles (surtout de type périphérique) ont tendance à être réduites en durée et par conséquent, subissent une RV très marquée (par rapport aux voyelles produites à débit lent et/ou en position accentuée), comme en anglais par exemple (cf. Peterson & Barney [2], Lindblom [1] en suédois, Stevens & House [3], Fourakis [4], Hillenbrand *et al.* [5], etc.). Pour Lindblom [1], ce sont essentiellement les différences de durée (et de contexte consonantique) et non de débit ou d'accentuation qui affectent la RV, en suédois : plus une voyelle est longue, plus la valeur cible est atteinte, et vice-versa. En revanche, Gay 1978 [6] et Pols & van Son [7] montrent que la non-accentuation et les différences de débit favorisent davantage la RV, lorsqu'ils comparent des voyelles accentuées produite à débit rapide et non-accentuées produites à débit lent, ayant par conséquent la même durée. Par ailleurs, Fourakis [4] a montré que le contexte environnant affecte largement les valeurs centrales de formants (et par conséquent la RV), en comparaison avec le débit de parole ou l'accentuation. Autrement dit, des voyelles accentuées ayant des durées variables et produites à débit normal en contexte CV subiront une forme de RV due majoritairement aux effets des consonnes adjacentes et non de la durée. En effet, la durée de la transition vers la voyelle en contexte CV est un invariant dynamique en production des occlusives voisées et non-voisées (≈ 50 ms), Kent & Moll [8]. Au-delà des 50 ms, c'est la durée de la partie stable de la voyelle qui pourrait influencer la RV, qui est favorisée par la coarticulation d'une voyelle avec les consonnes et/ou voyelles adjacentes. D'après Stevens & House [3], les voyelles produites dans un environnement "nul" (i.e. en contexte [hVd] ou en isolation) ne sont pas réduites, à l'inverse de celles présentes dans les autres contextes consonantiques (cf. Lindblom [1], Stevens & House [3], Öhman [9] Fourakis [4], Hillenbrand *et al.* [5], Al-Tamimi & Ferragne [10], etc.). Cependant, les effets du contexte consonantique sur les voyelles ont été étudiés d'un point de vue statique, c'est-à-dire en ne retenant que les valeurs centrales des formants vocaliques (Stevens & House [3], Fourakis [4]. Al-Tamimi & Ferragne [10]). D'un point de vue dynamique, l'étude de l'effet consonantique sur les voyelles en contexte CV a été établie par Lindblom [1], Hillenbrand *et al.* [5] et Fowler [11], entre autres et cela dès le début de la transition vers la voyelle, jusqu'à l'état stable, ou plus loin encore. Hillenbrand *et al.* [5] ont étudié ces effets dynamiques en caractérisant les changements spectraux inhérents aux voyelles. Des mesures acoustiques ont été effectuées à 20% et à 70% de la durée de la voyelle. Les résultats de leur analyse discriminante montrent que le taux de classification correcte est en moyenne 6,10% supérieur lorsque les valeurs de *f*0 et $F_{1\text{-}3}$ à 20% et à 70% sont prises en compte par rapport aux valeurs de *f*0 et $F_{1\text{-}3}$ à 50% (94,10% pour le premier test et 88% pour le second). La prise en compte de la durée améliore légèrement le taux d'identification (de 6,10% à 6,30%). Dans l'étude proposée par Lindblom [1], l'auteur a investigué la dynamique des formants vocaliques, et plus particulièrement *F*2, en utilisant les pentes de l'équation du locus (LE) en suédois (suivant la formule : $F2_{onset} = m*F2_{milieu} + b$, où m et b représente la pente et l'ordonnée à l'origine, respectivement). Ses résultats montrent que les pentes de LE varient en fonction du lieu d'articulation des consonnes allant de /g/ (0.95) > /b/ (0.69) > /d/ (0.28), et sont considérés depuis comme un bon indicateur du lieu d'articulation. Modarresi *et al.* [12] ont montré que les valeurs de pentes de LE ne varient pas entre /k/ et /g/ (mais varient en contexte bilabial et dental). Fowler [11] explique que LE sert à caractériser, à la fois, le lieu d'articulation et le degré de coarticulation entre les consonnes et les voyelles : une pente forte (m = 1) indique une coarticulation maximale entre consonnes et voyelles (i.e. une résistance minimale de la coarticulation), tandis qu'une pente faible (m = 0) indique l'absence de coarticulation entre consonnes et voyelles (i.e. une résistance maximale de la coarticulation). Le lien entre le degré de coarticulation et la RV peut être expliqué suivant la relation linéaire entre $F2_{onset}$ et $F2_{milieu}$: les modifications de valeurs de $F2_{milieu}$ affecteront celles de $F2_{onset}$ et par conséquent celles des pentes de LE.

La caractérisation des effets consonantiques sur les voyelles, la RV, d'un point de vue dynamique par les changements spectraux inhérents aux voyelles ou par les

pentes de LE, s'appuient sur des mesures "statiques" à 2 instants de la durée totale de la voyelle. Je propose, dans ce travail, une méthode de caractérisation de la RV d'un point de vue entièrement dynamique, c'est-à-dire, en représentant la transition par sa droite de régression donnant la pente formantique (PF). La transition part de l'onset et se termine au milieu de la voyelle. En ce sens, le calcul de la PF intègre totalement la durée et prends en compte ses variations.

2. MÉTHODOLOGIE

2.1. Langue, locuteurs et corpus

Trois systèmes vocaliques ont été comparés : l'arabe marocain de Casablanca avec 5 voyelles /i: ə a: u u:/ (Hamdi [13]), l'arabe jordanien d'Irbid avec 8 voyelles /i i: e: a a: o: u u:/ (Bani Yassin & Owens [14]) et le français avec 11 voyelles orales /i e ɛ a ɑ ɔ o u y ø œ/ (AM, AJ et FR, respectivement). 10 locuteurs hommes par système, âgés de 20 à 30 ans et ne présentant aucun trouble du langage ni au niveau articulatoire, ni auditif, ont produit une liste de mots ayant des structures syllabiques de type : C_1VC, C_1VCV, ou C_1VC_2VC. C_1 ou C_2 contient l'une des 3 consonnes *phonologiquement* communes aux 3 systèmes /b d k/ ; qui est suivie de la voyelle étudiée. Les items ont été présentés aléatoirement avec 5 répétitions par locuteur, à l'intérieur d'une phrase porteuse (le protocole expérimental a été adapté aux locuteurs arabophones par l'utilisation du système d'écriture de l'arabe standard, sans vocalisation et avec 2 listes différentes adaptées au lexique des deux dialectes). La tâche des locuteurs consistait à produire les voyelles dans des Mots, des Syllabes et en Isolation à débit moyen et en style non-marqué (ex. [b**o:**se ~ b**o:** ~ **o:**] = "bisou" en AJ). Les enregistrements ont été effectués dans une chambre insonorisée et numérisés directement sur un PC avec un taux d'échantillonnage de 22 kHz, 16 bits, mono. Au total, le corpus rassemble 2196 voyelles en AM, 3258 en AJ et 4800 en FR, segmentées manuellement.

2.2. Analyse et traitement de données

Dans cette étude, seules les données correspondant à la réalisation en Mot ont été utilisées, car cette forme se rapprocherait le plus de la situation de parole produite *normalement*, avec une production naturelle de la voyelle. Seules les 3 voyelles /i a u/ (voyelles longues accentuées en AM & AJ) ont été analysées du fait qu'elles sont communes aux 3 systèmes et non intrinsèquement réduites. Des analyses acoustiques fines des 2 premiers formants de chaque voyelle ont été effectuées avec le logiciel Praat en utilisant l'algorithme d'extraction de formants "Burg" (équivalent à une analyse LPC, auto-corrélation, avec 24 coefficients LPC) avec une fenêtre d'analyse gaussienne de 12,5 ms et un pas de déplacement de 5 ms. Les valeurs formantiques de l'onset ont été déterminées en suivant la méthode proposée par Al-Tamimi [15], et correspondent à la valeur distante de 5 ms du début de la transition vers la voyelle. Les valeurs formantiques, extraites toutes les 5 ms, ont été vérifiées manuellement afin de corriger les possibles erreurs d'extraction automatique. Ces valeurs ont été ensuite converties en Bark (suivant la formule proposée par Schroeder *et al.* [16]) pour procéder à une normalisation de données entre les locuteurs. Les valeurs de PF et de l'ordonnée à l'origine (OO) - m et b respectivement, dans la formule $y = m*x + b$ - sont obtenues par une analyse de régression linéaire. Le calcul a été effectué de l'onset vocalique jusqu'au milieu de la voyelle pour chaque mesure, où les valeurs de x représentent le temps et les valeurs y, les formants. Ainsi, pour chaque voyelle, une série de 2 valeurs de PF et d'OO représentant *F*1 et *F*2 est obtenue. Compte tenu des observations rappelées précédemment, on suppose que :

H1 : La RV sera affectée par le lieu d'articulation de la consonne adjacente, suivant l'évolution suivante : /k/ > /b/ > /d/, d'après les résultats obtenus de l'équation du locus, Lindblom [1],

H2 : Les valeurs de PF se seront pas les mêmes en fonction des voyelles : sur *F*1, /a/ aura les valeurs les plus élevées, /i/ et /u/, les plus basses, sur *F*2, /u/ aura les valeurs les plus élevées, les plus basses pour /i/, et /a/ se situera entre les deux. (Stevens & House [3]),

H3 : Les valeurs de PF et d'OO entre les 3 systèmes seront différentes, avec une RV moins importante en FR, et plus importante en AM, puisque les langues possédant plus de voyelles présentent une RV moins importante par rapport à celles possédant moins de voyelles, pour conserver la distinction entre elles, (Manuel [17]).

3. RÉSULTATS ET DISCUSSION

Deux types d'analyses statistiques ont été appliqués aux valeurs des PF et d'OO de *F*1 et de *F*2 afin de tester les influences sur la RV : 1) une MANOVA à 3 facteurs (langues, consonnes et voyelles), en utilisant NCSS et 2) 2 Analyses Discriminantes (AD) avec validation croisée ayant comme paramètres d'entrée : les valeurs de PF et d'OO de *F*1 et *F*2 en incluant et en excluant la durée, en utilisant SPSS. Les valeurs de PF sont positives ou négatives indiquant la direction de la transition (i.e. une PF négative indique une transition descendante vers la voyelle). En ne prenant en compte que les valeurs absolues de PF, on peut tester nos hypothèses. Des valeurs de PF élevées indiquent une RV moindre et vice versa, car les valeurs cibles se rapprochent de celles de l'état stable vocalique.

3.1. RV vs. Lieu d'Articulation des consonnes

Les résultats montrent que le lieu d'articulation affecte la RV (H 1). Les valeurs de PF sur l'axe *F*1 sont élevées pour /k/, basses pour /b/ et intermédiaires pour /d/ ($p < 10^{-6}$). Sur cet axe, la RV évolue de la façon suivante : /b/>/d/>/k/. Sur l'axe *F*2, les valeurs de PF sont élevées pour /d/, basses pour /b/ et intermédiaires pour /k/ ($p < 10^{-6}$). Sur cet axe, la RV évolue de la façon suivante : /b/>/k/>/d/ (figure 1). Les valeurs d'OO sont significativement différentes en fonction du lieu d'articulation de la façon suivante : /b/>/d/>/k/ sur *F*1 ($p < 10^{-6}$) et /d/>/k/>/b/ sur *F*2 ($p < 10^{-6}$), (figure 2). Les résultats de l'AD donnent un taux de classification des consonnes par langue significatif de 66,1% (53,6%, durée exclue) en AM (χ^2, $p < 10^{-6}$), de 44,5% (41,10%, durée exclue) en AJ (χ^2, $p < 10^{-6}$) et de 50,6% (43,20%, durée exclue) en FR (χ^2, $p < 10^{-6}$).

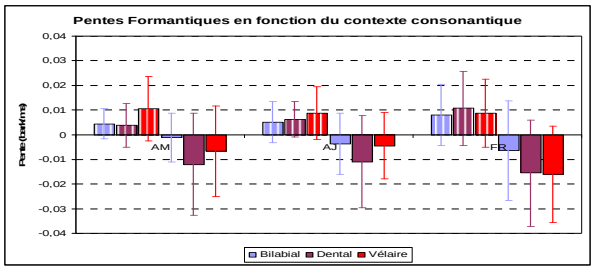


Figure 1 : PF de *F*1 (barres hachurée) et de *F*2 (barres pleines) en AM, AJ et FR en fonction des consonnes.

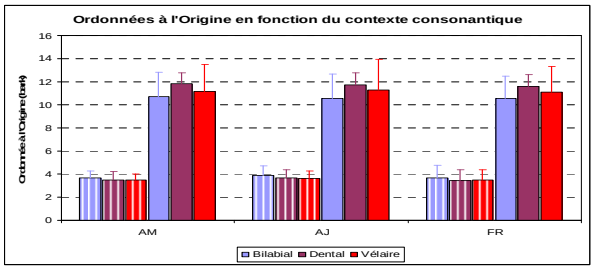


Figure 2 : OO de *F*1 (barres hachurée) et de *F*2 (barres pleines) en AM, AJ et FR en fonction des consonnes.

3.2. RV vs. Timbre de Voyelles

Les valeurs de PF, sur l'axe *F*1, sont élevées pour /a/, basses pour /i/ et intermédiaires pour /u/ ($p < 10^{-6}$). Sur l'axe *F*2, les valeurs de PF sont élevées pour /u/, basses pour /i/, et intermédiaires pour /a/ ($p < 10^{-6}$). Ceci montre que la voyelle /u/, sur l'axe *F*2, et la voyelle /a/, sur l'axe *F*1, subissent le maximum de RV (figure 3).

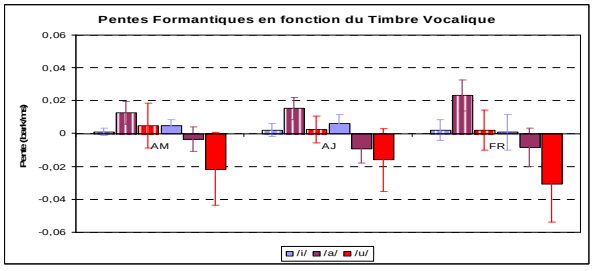


Figure 3 : PF de *F*1 (barres hachurée) et de *F*2 (barres pleines) en AM, AJ et FR en fonction des voyelles.

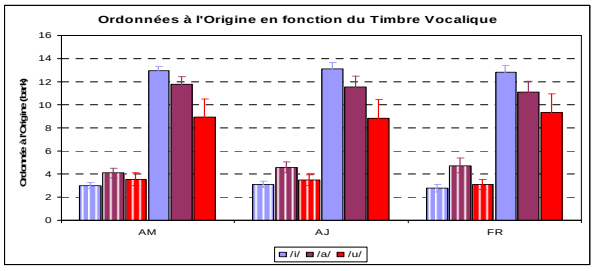


Figure 4 : OO de *F*1 (barres hachurée) et de *F*2 (barres pleines) en AM, AJ et FR en fonction des voyelles.

L'étude des effets consonantiques sur les voyelles montrent que, pour les 3 systèmes, la consonne /k/ affecte largement la voyelle /a/ sur l'axe *F*1 et la consonne /d/ réduit davantage les voyelles /i/ et /u/, sur l'axe *F*2, ($p < 10^{-6}$), expliquant ainsi une partie de la variabilité.

Les valeurs d'OO sont significativement différentes en fonction des voyelles de la façon suivante : /a/>/u/>/i/ pour *F*1 ($p< 10^{-6}$) et /i/>/a/>/u/ pour *F*2 ($p< 10^{-6}$), (figure 4). Les résultats de l'AD donnent un taux de classification des voyelles par langue très significatif de 99,8% (98,9%, durée exclue) en AM (χ^2, $p< 10^{-6}$), 98,4% (98,4%, durée exclue) en AJ (χ^2, $p< 10^{-6}$) et 98,4% (98,4%, durée exclue) en FR (χ^2, $p< 10^{-6}$).

3.3. RV vs. Différences Translinguistiques

Afin de tester cette hypothèse, des prédictions spécifiques doivent être formulées. En effet à durée identique, sur l'axe *F*1, une valeur de PF élevée pour /i/ et /u/ et basse pour /a/ indique une centralisation importante. Sur l'axe *F*2, plus la valeur de PF est basse, plus la voyelle est réduite (i.e. une centralisation pour /i/ et /u/ et antériorisée pour /a/). En observant les figures 3 & 4, toutes ces prédictions sont confirmées. En effet, on observe globalement une RV importante en AM et moins importante en FR, AJ étant intermédiaire ($p< 10^{-6}$). Néanmoins, comme le montre la table 1, il existe des différences importantes au niveau de la durée moyenne de chaque voyelle par langue (calculée de l'onset au milieu temporel). Ceci pourrait affecter le calcul de la PF, car au-delà de 50 ms (durée de transition), c'est la durée de l'état stable qui pourrait affecter la RV.

Table 1 : Durée moyenne et Écart-Type pour chaque voyelle (de l'onset au milieu temporel) par langue.

	/i/		/a/		/u/	
	moyenne	**écart-type**	**moyenne**	**écart-type**	**moyenne**	**écart-type**
AM	**127,50**	**32,37**	**115,07**	**29,92**	**78,61**	**32,38**
AJ	**110,77**	**26,72**	**129,63**	**21,60**	**90,41**	**28,58**
FR	**76,17**	**26,20**	**80,27**	**12,58**	**76,24**	**21,90**

Pour pallier ce problème, un second calcul de PF et d'OO a été effectué en normalisant le temps (i.e. en le situant dans un intervalle de 0 à 0,5, où 0 = le début de l'onset et 0,5 = le milieu temporel de la voyelle) afin d'éliminer les différences de durée entre les voyelles longues de l'arabe et les "brèves" du français. Les résultats obtenus montrent que pour /i/, les valeurs de PF sur l'axe *F*2 sont élevées en AJ et AM et basses en FR ($p< 10^{-6}$). Pour /a/, les valeurs de PF sur les deux axes *F*1 et *F*2 sont élevées en AJ, basses en AM et intermédiaires en FR ($p< 10^{-6}$). Pour /u/, les valeurs de PF sur l'axe *F*2 sont élevées en FR, basses en AM et intermédiaires en AJ ($p< 10^{-6}$) (figures 5). Ceci indique globalement une RV importante en AM, par rapport à AJ et FR, qui présente moins de RV, indiquant l'existence de différentes cibles vocaliques dans les 3 systèmes (cf. également Al-Tamimi & Ferragne [10]), surtout pour la voyelle /a/, où une RV importante est observée en FR par rapport à AJ. Les valeurs d'OO pour la voyelle /i/ sur l'axe *F*1 sont élevées en AJ, basses en FR et intermédiaires en AM ($p< 10^{-6}$). Sur l'axe *F*2, elles sont élevées en AJ par rapport à celles en AM et FR (qui se chevauchent). Pour la voyelle /a/ sur l'axe *F*1, les valeurs d'OO en FR sont élevées, basses en AM et intermédiaires en AJ ($p< 10^{-6}$). En revanche, elles sont élevées en AM, basses en FR et intermédiaires en AJ ($p< 10^{-6}$). Pour la voyelle /u/ sur l'axe *F*1, AM et AJ présentent les valeurs les plus élevées (qui se chevauchent) et FR, les plus basses ($p< 10^{-6}$). Sur l'axe *F*2, les valeurs sont élevées en

FR et basses en AJ et AM (les valeurs se chevauchent) ($p< 10^{-6}$), (figures 6).

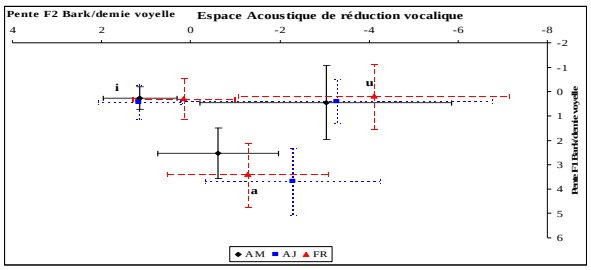


Figure 5 : Espace Acoustique des PF en AM, AJ et FR.

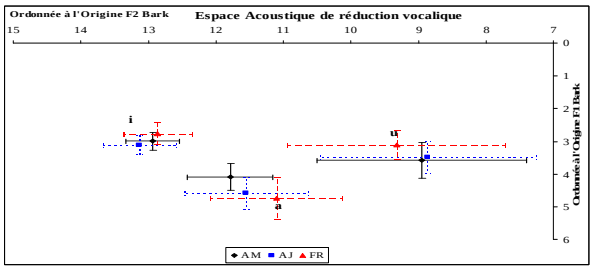


Figure 6 : Espace Acoustique des OO en AM, AJ et FR.

Les résultats de l'AD ont montré qu'il est possible de discriminer les 3 systèmes. En effet, l'analyse donne un taux de classification significatif de 36,7% pour les langues (χ^2, $p< 10^{-6}$) et 57,40% pour les voyelles (χ^2, $p< 10^{-6}$). Le taux de discrimination de chaque langue par catégorie vocalique est de : 61,1% pour /i/, 66,1% pour /a/ et 54,3% pour /u/, (χ^2, $p< 10^{-6}$).

4. CONCLUSION

Dans cette étude, une méthode de mesure de la RV basée sur les valeurs de PF et d'OO obtenues au moyen d'une régression linéaire a été proposée. Les résultats ont montré que les valeurs de PF et d'OO sont influencées par le lieu d'articulation des consonnes adjacentes (H 1), le timbre des voyelles (H 2) et par les différences translinguistiques de type "densité des systèmes" (H 3). L'utilisation de la durée et des valeurs de PF, d'OO, de *F*1 et de *F*2 lors d'une analyse discriminante a permis de séparer les 3 systèmes au niveau des consonnes et des voyelles. En excluant la durée du modèle, les taux d'identification restent approximativement les mêmes. La normalisation de la durée à travers les systèmes n'a pas changé les résultats : la RV est plus importante en AM, intermédiaire en AJ et plus basse en FR. Les résultats obtenus montrent qu'il est possible de reconnaître le lieu d'articulation de la consonne adjacente ou le timbre de la voyelle et cela dès le début de la transition. La prochaine étape consistera à analyser les autres voyelles et les autres contextes de réalisation, ainsi qu'à comparer avec de la parole spontanée. L'étude des attentes perceptives des auditeurs de chaque système permettra de mettre en évidence le rôle de ces indices dynamiques en perception.

5. REMERCIEMENT

Je remercie René Carré, François Pellegrino, Emmanuel Ferragne et Christelle Dodane pour leur aide précieuse.

BIBLIOGRAPHIE

[1] Lindblom, B. On vowel reduction, *Report #29, The Royal Ins. of Tech., Speech Transmission Laboratory, Stockholm, Sweden*, 1963.

[2] Peterson, G. & Barney, H. Control Methods Used in a Study of the Vowels, *Journal Acoustical Society of America*, Vol. 24: 175-184, 1952.

[3] Stevens, K. & House, A., Perturbation of vowel articulations by consonantal context: An acoustical study, *Journal of Speech and Hearing Researches*, Vol. 6: 111-128, 1963.

[4] Fourakis, M. Tempo, Stress and Vowel reduction in American English, *Journal of Acoustical Society of America*, Vol. 90 (4): 1816-1827, 1991.

[5] Hillenbrand, J., Clark, M. & Nearey, T. Effects of consonant environment on vowel formant patterns, *Journal of Acoustical Society of America*, Vol. 109 (2): 748-763, 2001.

[6] Gay, T. Effect of speaking rate on vowel formant movements, *Journal of Acoustical Society of America*, Vol. 63 (1): 223-230, 1978.

[7] Pols, L.; van Son, R. Acoustics and perception of dynamic vowel segments, *Speech Communication*, Vol. 13*:* 135-147, 1993.

[8] Kent, R. & Moll, K., Vocal-Tract Characteristics of Stop Cognates, *Journal of Acoustical Society of America*, Vol. 46 (6, part 2): 1549-1555, 1969.

[9] Öhman, S. Coarticulation in VCV Utterances: Spectrographic Measurements, *Journal of Acoustical Society of America*, Vol. 39: 151-168, 1966.

[10] Al-Tamimi, J. & Ferragne, E., Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. In *Proc. 9th EUROSPEECH*: 2465-2468, 2005.

[11] Fowler, C., Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation, *Perception & Psychophysics*, vol. 55**:** 597–610, 1994.

[12] Modarresi, G., Sussman, H., Lindblom, B., & Burlingame, E. Locus equation encoding of stop place: revisiting the voicing/VOT issue, *Journal of Phonetics*, vol. 33: 101-113, 2005.

[13] Hamdi, R. Étude phonologique et expérimentale de l'emphase en arabe marocain de Casablanca. *Thèse de Doctorat, Sciences du Langage: Lyon2*: 172, 1991.

[14] Bani-Yasin, R. & Owens, J. The Phonology of a Northern Jordanian Arabic Dialect, *Zeitschrift der Deutschen Morgenlandischen Gesellschaft*, Vol. 137(2): 297-331, 1987.

[15] Al-Tamimi, J., 2004, L'équation du locus comme mesure de la coarticulation VC et CV : Étude préliminaire en Arabe Dialectal Jordanien. In *Proc. of 25ème Journée d'Études sur la Parole*, pages 9-12, 2004.

[16] Schroeder, M., Atal, B., & Hall, J., Optimizing digital speech coders by exploiting masking properties of the human ear, *Journal of the Acoustical Society of America*, Vol. 66: 1647-1652, 1979.

[17] Manuel, S. Y. (1990). The Role of Contrast in Limiting Vowel-to-Vowel Coarticulation in Different Languages. *Journal of Acoustical Society of America*, Vol. 88(3): 1286-1298, 1990.

Session XIII

Production et perception

Mercredi 14 juin 2006 - 14h00 16h00

Estimation des dyspériodicités vocales dans la parole connectée dysphonique

A. Kacha [(1)], F. Grenez [(1)], J. Schoentgen [(1, 2)]

[(1)] Service Ondes et Signaux, Université Libre de Bruxelles, Bruxelles, Belgique
[(2)] Fond National de la Recherche Scientifique, Belgique
E-mail: akacha@ulb.ac.be

ABSTRACT

Acoustic analysis of connected speech is carried out by means of a generalized variogram to extract vocal dysperiocities. A segmental signal-to-dysperiodicity ratio is used to summarize the perceived degree of hoarseness. The corpora comprise four French sentences as well as vowels [a] produced by 22 male and female normophonic and dysphonic speakers. It is shown that the segmental signal-to-dysperiodicity ratio correlates better with perceptual scores of hoarseness than the global signal-to-dysperiodicity ratio. The perceptual scores are based on comparative judgments by six listeners of pairs of speech tokens.

1. INTRODUCTION

La présentation concerne la classification des voix dysphoniques. Les méthodes de classification basées sur les indices acoustiques sont populaires pour leur nature non invasive et permettent aux cliniciens de suivre l'évolution des patients et de quantifier le degré d'enrouement de la voix.

Plusieurs indices acoustiques sont utilisés pour caractériser la parole des locuteurs dysphoniques. Un nombre de ces indices reflète la déviation du signal de parole voisée par rapport à la périodicité parfaite. Les causes de ces dyspériodicités sont diverses : vibrations non modales des cordes vocales, bruit de modulation comprenant les variations cycle à cycle de la durée de cycle (jitter) et de l'amplitude (shimmer) dues aux perturbations externes, bruit additif dû à une turbulence excessive [7].

La plupart des indices acoustiques sont habituellement obtenus à partir de fragments stables extraits de voyelles soutenues. En effet, les voyelles soutenues sont faciles à analyser lorsque les attaques et les déclins sont exclus parce que les hypothèses de stationnarité et de cyclicité sont alors valables pour beaucoup de locuteurs. L'utilisation d'indices acoustiques obtenus à partir de voyelles soutenues est en fait justifiée par la faisabilité technique plutôt que par la pertinence clinique.

L'évaluation de la qualité de la voix est souvent basée sur la perception de la parole connectée. Par conséquent, on prévoit que les indices acoustiques obtenus à partir de la parole connectée soient mieux corrélés avec l'évaluation perceptuelle de la qualité de la voix. Plusieurs auteurs ont proposé d'extraire les indices acoustiques à partir de la parole connectée. En effet, la parole connectée contient les caractéristiques dynamiques de la source et du conduit vocal tels que les attaques et les déclins et les variations de la fréquence fondamentale et de l'amplitude [5] ce qui la rend plus informative que les voyelles soutenues.

Le nombre d'études sur la parole connectée est relativement faible par rapport aux études traitant les voyelles soutenues. Une revue des travaux publiés sur l'analyse de la parole connectée est donnée dans [1]. La plupart des méthodes d'estimation des dyspériodicities vocales, développées dans le cadre des voyelles soutenues, manquent de robustesse et de précision lorsqu'elles sont appliquées à la parole connectée ou à des voyelles avec attaque et déclin. Le manque de robustesse est une conséquence de l'hypothèse de stationnarité locale qui cesse d'être valable pour la parole connectée produite par des locuteurs enroués.

Dans [6] une approche basée sur un modèle prédictif a été proposée pour l'analyse des dyspériodicités vocales dans la parole connectée. Le modèle comprend un premier étage de prédiction à court terme conventionnel appliqué au signal de parole et un deuxième étage composé d'un prédicteur à long terme appliqué au résidu obtenu à la sortie du premier étage. Plus récemment, dans [1], un modèle de prédiction à long terme bilatérale appliqué directement au signal de parole a été proposé comme alternative. Le minimum des erreurs de prédiction à long terme dans les directions gauche et droite est retenu comme mesure des dyspériodicités, ce qui évite de comparer des fragments de signal à travers les limites phonétiques.

Les méthodes proposées dans [1] et [6] ne garantissent pas que les coefficients de prédiction à long terme soient toujours positifs, ce qui n'est pas cohérent avec la définition mathématique de la périodicité. Pour éviter ce problème, une méthode d'analyse basée sur le variogramme généralisé a été proposée [4].

L'indice acoustique conventionnel utilisé pour quantifier les dyspériodicités vocales dans le signal de parole est le rapport signal à dyspériodicité (RSD) global. La valeur numérique de l'indice global est principalement déterminée par les segments vocaliques dans la parole connectée [1]. Dans cette présentation, on se propose d'utiliser le rapport signal à dyspériodicité segmental (RSDSEG) comme indice acoustique. Il est obtenu par une reformulation locale de l'indice global. Les résultats montrent que le RSD segmental est plus performant que

le RSD global en termes de corrélation avec le degré d'enrouement perçu. Les scores de l'enrouement sont obtenus par une procédure basée sur la comparaison de paires de sons [3].

Le reste de la présentation est organisé comme suit. Dans la Section 2, le corpus utilisé dans l'expérience et les méthodes d'analyse sont présentés. Les résultats expérimentaux sont présentés et discutés dans la Section 3. La conclusion est donnée dans la Section 4.

2. METHODES

2.1. Corpus

Les données comprennent la voyelle [a], incluant attaque et déclin ainsi que quatre phrases produites par 22 locuteurs normophoniques ou dysphoniques (10 hommes et 12 femmes). Le corpus comprend 20 adultes (de 20 ans à 79 ans), un garçon âgé de 14 ans et une fille âgée de 10 ans. Cinq locuteurs sont normophoniques et les autres sont dysphoniques. Les phrases sont les suivantes : "Le garde a endigué l'abbé", "Bob m'avait guidé vers les digues", "Une poule a picoré ton cake" et "Ta tante a appâté une carpe". Ces phrases seront désignées par S1, S2, S3 et S4, respectivement. Elles ont la même structure grammaticale et le même nombre de syllabes. Les phrases S1 et S2 sont voisées par défaut alors que S3 et S4 comprennent des segments phonétiques voisés et non voisés.

Les signaux ont été enregistrés à une fréquence d'échantillonnage de 48 kHz, dans une cabine isolée, au moyen d'un enregistreur audio numérique (Sony TCD D8) et d'un microphone (AKG C41WL) au Département de Laryngologie d'un Hôpital Universitaire à Bruxelles, Belgique. Les enregistrements ont été transférés, par la suite, sur un disque dur d'ordinateur. Les intervalles de silence au début et à la fin des enregistrements ont été supprimés au moyen d'une segmentation manuelle.

2.2. Evaluation auditive

Une évaluation auditive basée sur des jugements comparatifs de paires de stimuli a été utilisée pour déterminer le degré d'enrouement de chaque échantillon (ou stimulus) du corpus composé de la voyelle [a] et des phrases S1 à S4 [3]. Il a été demandé aux auditeurs (ou juges) de comparer deux stimuli en terme du degré d'enrouement. L'objectif est d'hiérarchiser les stimuli, du moins enroué au plus enroué, au moyen de jugements comparatifs de toutes les paires de stimuli extraites de chaque ensemble homogène. La procédure d'évaluation est résumée comme suit :

1. La liste de toutes les paires de stimuli est formée sur la base de l'ensemble des enregistrements. Si N_e représente le nombre d'enregistrements (22 dans notre cas), le nombre de paires de stimuli est $N_e(N_e - 1)/2$.

2. Tous les scores des stimuli sont initialisés à zéro.

3. Une paire de stimuli choisie aléatoirement dans la liste est présentée à un auditeur qui doit désigner le stimulus le plus anormal de la paire. L'auditeur a aussi la possibilité de désigner les deux stimuli comme perceptivement identiques.

4. Le score total du stimulus désigné comme étant le plus anormal est augmenté d'une unité. Si les deux stimuli de la paire sont jugés égaux en termes du degré d'enrouement, leurs scores sont augmentés de 0.5 chacun.

5. Les étapes 3 et 4 sont répétées jusqu'à ce que toutes les paires de stimuli appartenant à une même session soient présentées.

6. Les stimuli sont alors caractérisés par leurs scores totaux. Le stimulus qui a été le plus souvent désigné comme étant le plus enroué aura le score le plus élevé tandis que le stimulus le moins enroué recevra le plus faible score.

La procédure est appliquée successivement à la voyelles [a] et aux phrases S1 à S4.

Les stimuli sont présentés via une interface audio numérique-analogique (Digidesign Mbox) et des écouteurs stéréo dynamiques (Sony MDR-7506). L'amplitude des sons est fixée par les auditeurs à un niveau confortable.

Le groupe de juges est composé de 6 auditeurs (une femme, cinq hommes) ayant tous une audition normale. Leurs ages varient de 24 ans à 57 ans. La tâche des auditeurs consiste à classer les stimuli sur la base du degré total de déviance de la voix. Chaque session d'audition est consacrée à un ensemble de 22 stimuli. Le nombre total de sessions est donc égal à 6 x 5 = 30. La même expérience a été répétée par cinq auditeurs après une période d'un jour au moins pour vérifier l'agrément intra-juges.

La moyenne des scores assignés par les six juges a ensuite été choisie comme une mesure subjective du degré d'enrouement perçu.

2.3. Variogramme généralisé

Pour un signal $x(n)$ périodique de période T_0, on peut écrire

$$x(n) = x(n - kT_0),\ k = \cdots, -2, -1, 0, 1, 2, \cdots \qquad (1)$$

Une mesure de l'écart par rapport à la périodicité sur un intervalle de longueur N fournit une indication sur le degré d'irrégularité du signal. Pour les signaux stationnaires, l'énergie des dyspériodicités peut être estimée par

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \left(x(n) - x(n-T) \right)^2 \right], \qquad (2)$$

avec $-T_{\max} \le T \le -T_{\min}$ et $T_{\min} \le T \le T_{\max}$.

L'expression entre crochets dans (2) est connue sous le nom de variogramme et est formellement équivalente à la

différence entre la trame d'analyse courante et une trame décalée de même longueur *N*. L'index temporel *n* positionne les échantillons du signal de parole à l'intérieur de la trame d'analyse. Les bornes T_{min} et T_{max} sont, en nombre d'échantillons, les cycles glottiques acceptables les plus courts et les plus longs. Ils son fixés à 2.5 ms et 20 ms, respectivement (50 Hz ≤ F0 ≤ 400 Hz). Pour les sons voisés, le délai *T* est interprété comme un multiple de la longueur du cycle glottique. Pour les sons non voisés, l'expression (2) demeure valide mais le délai *T* n'est pas interprétable en termes de longueur de cycle glottique.

L'amplitude du signal évolue d'une trame à la suivante à cause des attaques et des déclins, de l'intensité des segments et de l'accentuation. En introduisant un facteur de pondération *a* pour tenir compte de ces variations lentes de l'amplitude du signal, la définition (1) devient

$$x(n) = a\, x(n - kT_0),\ k = \cdots, -2, -1, 0, 1, 2, \cdots \qquad (3)$$

Selon cette définition, le variogramme généralisé prend alors la forme suivante

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \left(x(n) - a\, x(n-T) \right)^2 \right]. \qquad (4)$$

Le gain *a* doit être positif. Il est défini de manière à garantir des énergies identiques dans la fenêtre d'analyse courante et la fenêtre décalée

$$a = \sqrt{\frac{E}{E_T}}, \qquad (5)$$

où *E* et E_T sont les énergies des trames d'analyse courante et décalée,

$$E = \sum_{n=0}^{N-1} x^2(n),\ E_T = \sum_{n=0}^{N-1} x^2(n-T).$$

La longueur de la trame d'analyse *N* et la longueur du décalage sont fixées à 2.5 ms. Ce choix permet de garantir que chaque fragment du signal soit inclus exactement une seule fois dans l'analyse. La valeur instantanée de la dyspériodicité est estimée comme suit :

$$e(n) = x(n) - a\, x(n - T_{opt}),\quad 0 \le n \le N-1 \qquad (6)$$

où T_{opt} est le délai qui minimise le variogramme généralisé (4) pour la position courante de la trame d'analyse. L'analyse est effectuée dans les directions gauche et droite et, par conséquent, le délai T_{opt} peut prendre des valeurs positives ou négatives.

L'approximation du délai optimal par un nombre entier de périodes d'échantillonnage introduit un bruit de quantification. L'effet de ce bruit de quantification est réduit en suréchantillonnant le signal d'un facteur 8.

2.4. Indices acoustiques global et segmental

L'indice acoustique conventionnel utilisé pour quantifier les dyspériodicités vocales dans le signal de parole est le rapport signal à dyspériodicité global exprimé par [1]

$$RSD = 10 \log \left[\frac{\sum_{n=0}^{L-1} x^2(n)}{\sum_{n=0}^{L-1} e^2(n)} \right], \qquad (7)$$

où *e*(*n*) est la dyspériodicité instantannée estimée selon (6) et *L* est le nombre d'échantillons dans l'intervalle total d'analyse

Le rapport signal à dyspériodicité segmental (RSDSEG) est connu comme un bon estimateur de la qualité de la parole dans le contexte du codage [2]. Le RSDSEG est obtenu par une reformulation locale du RSD global en calculant le RSD sur des segments courts de l'intervalle d'analyse et en prenant la moyenne de tous les RSD. On prévoit que le RSDSEG d'une production sera mieux corréle avec le degré d'enrouement que le RSD global. En effet, la valeur du RSDSEG est obtenue en appliquant la fonction logarihmique avant de moyenner sur l'ensembles des mesures locales, ce qui permet de donner une plus forte pondération aux segments bruités de faibles niveaux qui sont peu pondérés dans le calcul du RSD global. Par conséquent, les segments de grande amplitude et peu bruités ne masquent pas numériquement la contribution des segments bruités de faible amplitude.

Pour une production donnée, l'intervalle d'analyse est divisé en *K* segments de longueurs *M* et le RSDSEG est calculé comme suit :

$$RSDSEG = \frac{10}{K} \sum_{k=0}^{K-1} \log \frac{\sum_{n=Mk}^{Mk+M-1} x^2(n)}{\sum_{n=Mk}^{Mk+M-1} e^2(n)}. \qquad (8)$$

3. RESULTATS ET DISCUSSION

Les scores de l'enrouement perçu ont été déterminés par six auditeurs. Les scores moyens dépendent légèrement du type de production. Ils varient entre 2.2 et 7.5 pour les locuteurs normophoniques et entre 3.2 et 20.4 pour les locuteurs dysphoniques sur une échelle allant de 0 à 21.

Le RSDSEG a été calculé pour différentes valeurs de la longueur *M* des segments. Les résultat de l'analyse de corrélation entre le RSDSEG et les scores moyens de l'enrouement perçu sont donnés dans le tableau 1 pour les différentes productions (voyelles [a] et phrases S1 à S4). La corrélation dépend légèrement de la longueur du segment et se stabilise à 5 ms. Par la suite, la longueur des segments dans le calcul du RSDSEG a été fixée à cette valeur. Les valeurs du RSDSEG varient de 17.8 dB à 23.8

dB pour les locuteurs normophoniques et de 5.5 dB à 23.8 dB pour les locuteurs dysphoniques.

Le coefficient de corrélation de Pearson des scores moyens de l'enrouement avec les rapports signal à dyspériodicité global et segmental a été calculé et les résultats pour la voyelle [a] et les phrases S1 et S4 sont donnés dans le tableau 2. L'hypothèse nulle ($\rho_P = 0$) a été rejetée pour toutes les entrées du tableau (test unidirectionnel, $\rho_{crit} = 0.36$, $p < 0.05$). L'inspection des résultats du tableau 2 montre que le RSD segmental est plus fortement corrélé que le RSD global pour les phrases S1 à S3. Ceci s'explique par le fait que dans le jugement des sons, les auditeurs sont influencés par les segments bruités quoiqu'ils soient courts et par les fragments vocaliques caractérisés par un grand rapport signal à dyspériocité. Le RSD segmental en fonction du degré d'enrouement correspondant pour la phrase S1 est représenté sur la figure 1.

Le fait que, pour la voyelle, la performance du RSDSEG ne soit pas améliorée, en termes de corrélation avec le degré d'enrouement, par rapport à celle du RSD global est attendu. En effet, les dyspériodicités sont également distribuées dans les sons de parole soutenue. On observe aussi que la performance du RSD n'est pas améliorée pour la phrase S4. Une explication de cette constatation serait que l'évaluation perceptive est moins fiable pour la phrase S4. Les auditeurs ont en effet rapporté que la phrase S4 était relativement difficile à évaluer parce que les intervalles voisés apparaissaient très courts.

Tableau 1 : Coefficients de corrélation de Pearson entre les valeurs du RSD segmental et les scores moyens de l'enrouement pour les voyelle [a] et les phrases S1 à S4. La longueur des segments est indiquée dans la colonne de gauche.

	[a]	S1	S2	S3	S4
30 ms	– 0.71	– 0.84	– 0.80	– 0.78	– 0.65
20 ms	– 0.71	– 0.84	– 0.80	– 0.79	– 0.67
10 ms	– 0.71	– 0.85	– 0.81	– 0.81	– 0.68
5 ms	– 0.70	– 0.86	– 0.81	– 0.81	– 0.70
2.5 ms	– 0.70	– 0.86	– 0.81	– 0.82	– 0.70

Tableau 2: Coefficients de corrélation de Pearson des rapports signal à dyspériodicité global et segmental avec les scores d'enrouement. La longueur des segments est fixée à 5 ms pour le calcul du RSD segmental.

	RSD global	RSD segmental
[a]	– 0.73	– 0.70
S1	– 0.72	– 0.86
S2	– 0.72	– 0.81
S3	– 0.70	– 0.81
S4	– 0.69	– 0.70

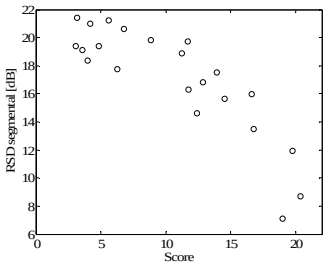


Figure 1 : RSDSEG en fonction des scores d'enrouement associés à la phrase S1 (22 locuteurs) pour une longueur de segment de 5 ms.

4. CONCLUSION

Dans cette présentation, une méthode d'estimation des dyspériodicités vocales dans la parole connectée, basée sur le variogramme généralisé, a été proposée. Le rapport signal à dyspériodicité segmental a été utilisé pour quantifier les dyspériodicités vocales. La performance en terme de corrélation avec le degré d'enrouement perçu a été comparée avec celle du rapport signal à dyspériodicité global conventionnellement utilisé dans l'analyse de la parole dysphonique. Les résultats montrent que l'indice segmental est mieux corrélé avec le degré d'enrouement que l'indice global.

BIBLIOGRAPHIE

[1] F. Bettens, F. Grenez and J. Schoentgen. Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis. *J. Acoust. Soc. Am.*, 117: 328-337, 2005.

[2] N.S. Jayant and P. Noll. Digital coding of waveforms :principles and applications to speech and video, Prentice-Hall, Englewood Cliffs, 1984.

[3] A. Kacha, F. Grenez and J. Schoentgen. Voice quality assessment by means of comparative judgments of speech tokens. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 1, pages 1733-1736, 2005.

[4] A. Kacha, F. Grenez, J. Schoentgen and K. Benmahammed. Dysphonic speech analysis using generalized variogram. In *Proc. Intl. Conf. on Acoustic Speech and Signal Processing*, volume 1, pages 917-920, 2005.

[5] F. Klingholtz. Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels. *J. Acoust. Soc. Am.*, 87: 2218-2224, 1990.

[6] Y. Qi, R.E. Hillman and C. Milstein. The estimation of signal-to-noise ratio in continuous speech of disordered voices. *J. Acoust. Soc. Am.*, 105: 2532-2535, 1999.

[7] J. Schoentgen. Spectral models of additive and modulation noise in speech and phonatory excitation signals. *J. Acoust. Soc. Am.*, 113: 553-562, 2003.

L'intégration bimodale de l'anticipation du flux vocalique dans le flux consonantique

Emilie Troille & Marie-Agnès Cathiard

Institut de la Communication Parlée, Université Stendhal-INPG, UMR 5009,
38040 Grenoble cedex 9, France
Mail : troille@icp.inpg.fr ; cathiard@icp.inpg.fr

ABSTRACT

It is well known that speech can be seen before it is heard: this has been repeatedly shown for the vowel rounding anticipatory gesture leading the sound (Cathiard [6]). In this study, the perception of French vowel [y] anticipatory coarticulation was tested throughout a voiced fricative consonant [z] with a gating paradigm. It was found that vowel auditory information, as carried by the noise of the fricative, was ahead of visual and even audiovisual information. Hence the time course of bimodal information in speech cannot be considered to display the same pattern whatever the timing of the coordination of speech gestures. As concerns vowel information only, consonantal coarticulation can carry earlier auditory information than the vowel itself, this depending of the structure of the stimulus. In our fricative-vowel case, it was obvious that the vowel building movement was audible throughout the fricative noise, whereas the changes in formant grouping occured later.

1. INTRODUCTION

D'importantes questions demeurent concernant l'intégration des informations visuelles et auditives en dépit des apports significatifs à ce champ de recherche depuis une vingtaine d'années (cf. Schwartz [1] pour une revue). Dire que la parole est avant tout bimodale peut sembler aujourd'hui être une affirmation triviale, même si le poids de la contribution de l'une et l'autre des modalités n'est pas encore complètement déterminé. Audition et vision jouent-elles toujours en complémentarité, comme on a pu le conclure à partir de paradigmes contrastant une présentation audiovisuelle à une information monomodale (avec une information auditive soit dégradée par du bruit ou bien parfaitement audible mais sémantiquement difficile à comprendre) ?

Afin de mettre en évidence la contribution des informations auditives et visuelles présentées seules ou associées, nous avons retenu un cas classique en parole, celui de l'anticipation vocalique d'arrondissement, qui nous offre une situation dans laquelle l'audio pourra être testé sans dégradation par le bruit, tout en évitant les effets plafonds habituellement rencontrés dans les expériences qui ne dégradent précisément pas l'information auditive.

L'anticipation d'arrondissement vocalique a fait l'objet de nombreuses études tant au niveau articulatoire qu'au niveau perceptif. Il est ainsi établi que ce geste vocalique peut débuter plusieurs consonnes avant la voyelle arrondie cible (pour une synthèse concernant l'anglais cf. Perkell [2] ; pour le français cf. Abry et Lallouache [3]). La récupération auditive de cette anticipation a été mise en évidence en français par Benguerel et Adelman [4], et plus récemment par Hecker et al. [5], cette dernière étude montrant que l'anticipation était maximale avec des consonnes intervocaliques fricatives plutôt qu'occlusives.

L'identification visuelle de l'anticipation labiale a été extensivement étudiée par Cathiard [6] au cours de pauses acoustiques silencieuses. Les séquences UHI et IHI étaient insérées dans la phrase « tu dis : UHI ise ? » [tydi#uiiz], avec une pause (#) longue de 460 ms. Des mesures articulatoires montraient que le geste d'arrondissement s'établissait au cours de la pause silencieuse de telle sorte que la position en protrusion de la lèvre supérieure et en constriction, caractéristique de la voyelle arrondie [y] à venir, soit établie dès le début acoustique de celle-ci. Les tests d'identification indiquaient une anticipation de la perception visuelle du [y] pouvant aller jusqu'à plus de 200 ms avant le début acoustique de la voyelle.

Les études ayant directement comparé les modalités auditive et visuelle sont peu nombreuses. Escudier et al. [7], testant entre autres des séquences [zizy], concluaient à une avance du visuel sur l'audio (de 40 à 60 ms), qui est vraisemblablement, de notre point de vue, à relier à un geste de constriction pour [y] extrêmement précoce, soit dès la deuxième partie de la voyelle [i]. Dans une étude plus récente Roy [8] comparait la perception auditive et visuelle de sujets malentendants et normo-entendants pour des séquences [iCny] (Cn représentant un nombre variable de consonnes). En modalité auditive, la perception de la voyelle arrondie se situe au moment du relâchement consonantique lorsque l'intervalle n'est composé que d'occlusives, et à partir de l'inflexion de la limite inférieure du bruit de friction lorsque l'intervalle contient une fricative. Pour ce dernier cas, Roy constate que les sujets normo-entendants sont plus performants en modalité auditive qu'en modalité visuelle pour percevoir le trait d'arrondissement, mais sans que l'on puisse estimer précisément l'avance.

Dans cette étude, nous nous proposons de tester l'établissement de l'information vocalique à travers le flux consonantique sous 3 conditions de présentation : auditive, visuelle et audiovisuelle.

2. EXPERIENCE PERCEPTIVE

2.1. Enregistrement

Nous avons enregistré audiovisuellement, à 25 images/seconde, de face et de profil, un locuteur masculin français, prononçant en ordre aléatoire 12 répétitions des 2 séquences « t'as dit ZIZU ze ? » et « t'as dit ZIZI ze ? ».

Ces phrases nous permettront d'explorer la transition de la voyelle [i] vers la voyelle [y] avec une consonne [z] intervocalique. Il est bien connu que cette fricative est perméable aux effets de coarticulation (Öhman [9]).

L'enregistrement a été réalisé en chambre sourde, au moyen du poste « Visage-Parole » de l'ICP (Lallouache [10]). Le son est échantillonné à 22050 Hz. Un maquillage en bleu des lèvres du sujet permettra par la suite l'incrustation d'un noir saturé, à l'aide d'un Chroma-Key, afin de permettre la détection des paramètres labiaux.

2.2. Analyse des données

Les images numérisées ont été détramées, afin d'obtenir une image toutes les 20 ms. La détection de 2 paramètres articulatoires permettant de caractériser le geste d'arrondissement, soient l'aire intérolabiale (S) et la protrusion de la lèvre supérieure (P1), est réalisée à l'aide du logiciel Tacle (Audouy [11]).

Nous avons réalisé, au niveau acoustique, un suivi de formants (F1 à F4) pour toutes nos séquences afin de trouver une réalisation [zizy] et [zizi] qui soient le plus semblables possibles, au moins en ce qui concerne le début de la transition [zi…] (fig. 1). Nous avons également suivi le bruit de friction de la consonne dans les deux transitions par une analyse LPC (en utilisant une fenêtre d'analyse large de 0.04 s de manière à privilégier l'effet de coarticulation avec la voyelle suivante, cf. Munson [12]). Dans la transition [zizy] retenue, on peut observer un abaissement du bruit de friction de la consonne [z] qui passe de 5653 Hz à 3293 Hz (alors qu'il reste stable dans la transition contrôle [zizi] autour d'une valeur moyenne de 6067 Hz), abaissement qui est à relier à la diminution de l'aire aux lèvres au cours du [z] par anticipation de l'arrondissement du [y] (fig. 2); on remarque aussi dès la fin du [z], une baisse du 3ème formant pour [zizy] jusqu'à 2400 Hz tandis qu'il reste aux alentours de 3000 Hz dans [zizi].

2.3. Montage des tests

Nous adoptons une technique de dévoilement progressif ou « gating » largement utilisée en perception de parole, afin de tester l'identification auditive, visuelle et audiovisuelle de l'arrondissement vocalique à travers la consonne. Dans les 3 conditions, les séquences commencent toutes au début de « t'as dit » et se terminent aux différents temps de la zone de *gating*. Ce domaine d'exploration démarre 40 ms avant la fin du [i] (soit, pour repère, à 1560 ms du début de la phrase) et se termine à la fin de la consonne [z] (soit au temps de 1720 ms). Le pas du gating étant de 20 ms, nous obtenons 9 séquences tronquées, que nous répétons 10 fois pour ZIZU, auxquelles nous ajoutons 3 séquences pour ZIZI (les séquences tronquées respectivement à 1660, 1680 et 1720 ms, répétées 5 fois ; ces séquences permettent au sujet d'entendre et/ou voir de vraies séquences ZIZI sans pour autant alourdir la passation en proposant les 9 séquences tronquées de ZIZI).

Le découpage audio du domaine perceptif toutes les 20 ms est réalisé sous Praat, à l'aide d'un Script. Pour le test audiovisuel, nous ajoutons sous Adobe Première Pro 1.5., les images détramées. Pour le test visuel, nous montons les images avec seulement le début de la séquence audio, soit « t'as dit », pour servir d'amorce attentionnelle au sujet. Les tests sont ensuite insérés dans le logiciel Multimédia Toolbook., qui permet le tri aléatoire des séquences présentées aux sujets.

2.4. Sujets

26 sujets de langue maternelle française, sans déficience ni auditive ni visuelle (ou corrigée) sont testés. Aucun n'est familiarisé avec ce type de test, ou n'a de connaissance particulière en lecture labiale. Le groupe est composé de 2 hommes et de 24 femmes, de 20 à 25 ans (21 ans et 3 mois en moyenne). Ils commencent tous par le test audiovisuel, suivi du test audio, et terminent par le test visuel : nous avons retenu cet ordre car nous souhaitions recueillir en premier l'identification audiovisuelle sans qu'il y ait eu d'entraînement préalable en condition unimodale. La tâche des sujets consiste à d'identifier la voyelle finale [i] ou [y] de chaque séquence tronquée.

3. RESULTATS

3.1. Résultats par modalités

Nous présentons les courbes d'identification, pour tous les sujets confondus, obtenues pour [zizy].

(1) L'identification auditive augmente rapidement et de façon très nette entre 1620 et 1640 ms (fig. 3). La voyelle [y] est identifiée à plus de 80% à 1640 ms, soit clairement dans la consonne [z] qui précède, puisque le début effectif de cette voyelle est situé à notre dernier temps de gating, soit à 1720 ms. Les identifications individuelles sont relativement regroupées autour de la courbe moyenne, la frontière à 50% d'identification [y] variant de 1610 à 1640 ms.

(2) L'identification visuelle évolue très lentement (fig. 4). La frontière à 50% correspond au temps 1680 et l'identification du [y] à plus de 80% n'est atteinte que 20 ms avant le début acoustique de la voyelle. On observe un empan de variation des identifications individuelles à 50% de 1640 à 1710 ms, soit 70 ms, ce qui reflète un comportement perceptif visuel variable selon le sujet.

(3) Les résultats du test audiovisuel montrent que la voyelle arrondie est perçue à plus de 80 % à 1680 ms en condition bimodale (fig. 5). A 1660 ms, [y] est déjà perçu à 76 %. La dispersion autour de la courbe moyennée est intermédiaire par rapport à celles observées pour les 2 conditions unimodales, avec un empan à 50% d'environ 50 ms (de 1620 à 1670 ms).

3.2. Comparaison des modalités

La figure 6 présente les courbes d'identification moyennées pour les 3 conditions (auditive A, audiovisuelle AV et visuelle V). L'identification auditive de la voyelle [y]

apparaît être la plus précoce. Elle est suivie de l'identification audiovisuelle, puis de l'identification visuelle. L'analyse des données individuelles indique que cet ordre est observé chez 24 sujets sur 26.

Nous avons réalisé une analyse Probit (Finney [13]) pour extraire les frontières et les pentes pour chacun des sujets dans les trois conditions. Nous réalisons, sur les frontières ainsi que sur les pentes, une analyse de la variance à un facteur intra-sujets, soit la condition de présentation (A, AV et V). En ce qui concerne les frontières, après correction de l'homogénéité des variance (Greenhouse-Geisser), la condition de présentation a un effet significatif : $F(1,5,38)=172.42$ avec $p=0.000$, avec A > AV > V. Ainsi, la frontière moyenne à 50% est à 1629 ms en audio, 1649 ms en audiovisuel et 1673 ms en visuel.

L'étude des pentes montre également un effet de la condition : $F(2,50)=7.451$; $p<0.001$, avec cette fois l'homogénéité des variances vérifiée sans correction (test de Sphéricité de Mauchly). On ne trouve pas de différence de pente entre la condition auditive et la condition audiovisuelle ($F<1$). On peut donc associer les pentes de l'audio et de l'audiovisuel pour les comparer au visuel. On obtient cette fois une différence significative : $F(1,25)=11.18$; $p<0.003$. La bascule d'identification est ainsi comparable en audio et en audiovisuel, mais plus lente en identification visuelle.

3.3. Relation perception-production

La comparaison des identifications auditive et audiovisuelle au suivi du bruit de friction de [z] (fig. 7) montre que la perception de la voyelle [y] augmente dès que le mouvement de bruit de friction commence à descendre. On note un retard de 20 ms de l'identification audiovisuelle sur l'identification auditive, [y] étant perçu auditivement lorsque la fréquence de [z] passe en dessous de 4000 Hz contre 4500 Hz en audiovisuel.

En vue de face, le paramètre visuel le plus facile à suivre étant l'aire aux lèvres (fig. 8), nous le comparons au suivi de l'identification visuelle et audiovisuelle [y]. [y] est ainsi identifié lorsque l'aire aux lèvres passe en dessous de 1 cm² en condition audiovisuelle : l'identification passe de 23% lorsque S est à 1,12 cm² au temps 1640 ms à 76% lorsqu'elle est à 0,85 cm² au temps 1660 ms. En revanche, la bascule est plus tardive et plus lente pour l'identification visuelle : elle s'établit au temps 1680 ms, lorsque l'aire au lèvres est inférieure à 0,62 cm². Il semble qu'en vision seule, les sujets suivent plutôt l'évolution de la protrusion de la lèvre supérieure (fig. 9), comme s'ils attendaient une forme aux lèvres à la fois bien arrondie et protruse pour être sûrs de l'identification de la voyelle.

En résumé, en modalité auditive, les sujets utilisent au maximum les informations acoustiques pour identifier [y] en suivant l'abaissement du bruit de friction de la consonne [z] comme un indice d'arrondissement de la voyelle à venir, tandis qu'en condition visuelle, les indices articulatoires ne leur permettent pas d'être aussi précoces. L'abaissement du bruit de friction est bien dû au mouvement de constriction labiale du [y], et il semble qu'un petit mouvement audio suffise aux sujets pour commencer à identifier [y], alors qu'il leur faut attendre un seuil de constriction parfois inférieur au cm² pour identifier visuellement la voyelle. Dans la condition audiovisuelle, il semble que les sujets se soient laissés influencer par les deux modalités à la fois.

4. DISCUSSION

Notre objectif était de tester la perception uni- et bimodale du flux vocalique en présence d'un flux consonantique continu.

(1) Les résultas en audio seul mettent en évidence une perception précoce de la voyelle arrondie dans la consonne qui la précède, puisque le [y] est perçu dès la fin du premier tiers de la consonne (avance de près de 90 ms dans une consonne de 120 ms). L'identification audiovisuelle suit avec un retard de 20 ms, tandis que la perception visuelle est la moins précoce, accusant un retard de 44 ms par rapport à l'audio (ce qui la place néanmoins encore à 47 ms du début acoustique de la voyelle). Concernant l'avance de la vision sur l'audition défendue par Cathiard [7], rappelons qu'elle se produisait pour une voyelle produite à l'initiale, sans consonne, et avec une structure prosodique démarcative (pause silencieuse) : dans ce cas précis où le flux vocalique est perceptible isolément, la vision est forcément gagnante puisque l'audio est délivré de manière naturelle beaucoup plus tardivement. Dans le cas de notre séquence [zizy], les informations articulatoires (constriction labiale) et acoustiques (bruit de friction) sont parfaitement synchrones ; de plus, le flux vocalique se trouve en présence d'un flux consonantique. Dans ce cas précis, l'auditeur piste la voyelle aussitôt que possible à travers la consonne en s'appuyant sur l'évolution des zones d'énergie depuis l'abaissement du bruit de friction de la consonne jusqu'au 3ème formant de la voyelle.

(2) Nous démontrons également que l'identification audiovisuelle se situe temporellement entre les deux présentations unimodales. On pourrait rapprocher ce résultat de certaines interprétations de l'illusion McGurk qui défendent un percept audiovisuel juste intermédiaire entre les stimuli audio et visuel. Mais n'oublions pas que nous ne sommes pas, dans cette expérience, dans le cas d'une information conflictuelle, mais dans le cas d'un flux congruent, qui s'avère récupérable plus précocement dans une modalité que dans l'autre. On pourrait aussi penser que notre résultat en audiovisuel contredit la règle de la supériorité de l'information bimodale sur l'information unimodale : cette règle est en réalité valable en situation bruitée, ce qui n'est pas non plus notre cas. Nous sommes donc en présence d'une situation où, avec un son parfaitement audible et un flux visuel non contradictoire, on observe néanmoins un certain retard (20 ms en moyenne) de la perception audiovisuelle sur la perception auditive.

En conclusion, nous avons pu tester l'établissement de la perception de l'anticipation vocalique : l'information auditive vocalique, portée par le bruit de la fricative, est perçue en avance sur l'information visuelle et même sur la

perception audiovisuelle. Ainsi, l'évolution temporelle de l'information bimodale en parole dépend fortement du timing de la coordination des gestes articulatoires. En ce qui concerne l'information vocalique seule, la coarticulation consonantique peut la délivrer auditivement très en avance de la voyelle. Dans le cas étudié ici, le mouvement d'établissement de la voyelle est audible dans le bruit de friction de la fricative alors que les changements dans les groupements formantiques se produisent plus tardivement (cf. fig. 1).

BIBLIOGRAPHIE

[1] J.-L. Schwartz. La parole multisensorielle : Plaidoyer, problèmes, perspective. *XXVème Journées d'Etudes sur la Parole*, 11-18, Fès, Maroc, 19-21 avril 2004.

[2] J. S. Perkell. Testing theories of speech production : implications of some detailed analyses of variable articulatory data. In W.J. Hardcastle & A. Marchal (Eds), *Speech Production and Speech Modelling*, pp. 263-288, K.A.P., London, 1989.

[3] C. Abry. & T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur, données sur l'arrondissement en français. *Bulletin de la Communication Parlée,3*, 85-99, 1995.

[4] A.-P. Benguerel & S. Adelman. Perception of Coarticulated Lip Rounding. *Phonetica*, *33*, 113-126, 1976.

[5] V. Hecker, B. Vaxelaire., M.-A. Cathiard, C. Savariaux & R. Sock. How lip protrusion expansion influences auditory perceptual extent. Probing into the Movement expansion Model. In C.Cavé, I. Guaïtella & S. Santi (Eds.), *Oralité et Gestualité : Interactions et comportements multimodaux dans la communication*, 450-456, L'Harmattan, Paris, 2001.

[6] M.-A. Cathiard. La perception visuelle de l'anticipation des gestes vocaliques : cohérence des évènements audibles et visibles dans le flux de la parole. Thèse de Psychologie Cognitive, Grenoble, 1994.

[7] P. Escudier, C. Benoit & T. Lallouache. Identification visuelle de stimuli associés à l'opposition /i/-/y/ : étude statistique. *Actes du Premier Congrès d'Acoustique*, Lyon, 10-13 Avril, *Suppl. au Journal de Physique*, *2*, 541-544, 1990.

[8] J.-P. Roy. Etude de la perception des gestes anticipatoires d'arrondissement par les sourds et les malentendants. Thèse de Sciences du Langage, Strasbourg, 2004.

[9] S. E. G. Öhmann. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, *41(2)*, 310- 320, 1967.

[10] M.-T. Lallouache. Un poste "Visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres. Thèse de l'ENSERG, Spécialité: Signal Image Parole, Grenoble, 1991.

[11] M. Audouy. Logiciel de traitement d'images vidéo pour la détermination de mouvements des lèvres. Projet de fin d'études, option génie logiciel, ENSIMA Grenoble, 2000.

[12] B. Munson. Variability in /s/ Production in Children and Adults : Evidence from Dynamic Measures of Spectral Mean. *Journal of Speech, Language and Hearing Research, 47(1)*, 58-69, 2004.

[13] D.-J. Finney. *Probit Analysis*. Cambridge University Press, (3ème édition) 1971.

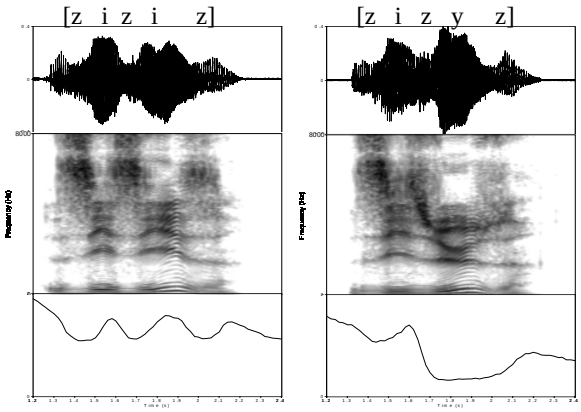


Fig. 1 : signal acoustique, spectrogramme (échelle de fréquence de 0 à 8KHz) et aire aux lèvres (de 0 à 2 cm²) pour ZIZI (à gauche) et ZIZU (à droite).

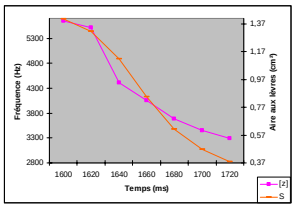


Fig. 2 : suivis de la fréquence du bruit de friction et de l'aire aux lèvres.

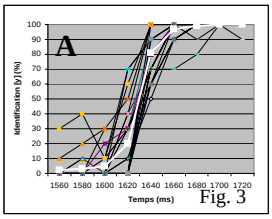


Fig. 3

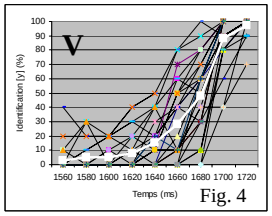


Fig. 4

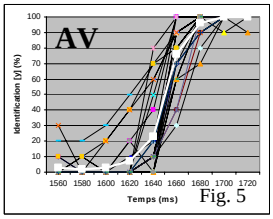


Fig. 5

Fig. 3, 4 et 5 : comparaison des courbes auditive, visuelle et audiovisuelle globales à la distribution des courbes individuelles.

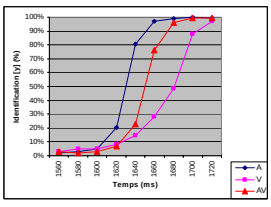


Fig. 6 : comparaison des courbes d'identification moyennées auditive, visuelle et audiovisuelle.

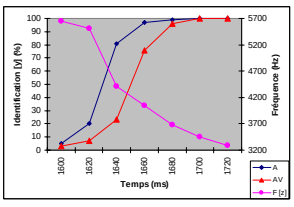


Fig. 7 : comparaison des courbes d'identification moyennées auditive et audiovisuelle au suivi du bruit de friction de [z].

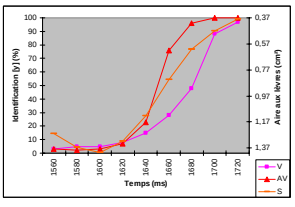


Fig. 8 : comparaison de l'aire aux lèvres (échelle inversée) aux courbes d'identification visuelle et audiovisuelle.

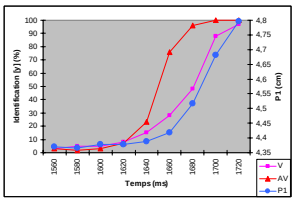


Fig. 9 : comparaison de la protrusion aux courbes d'identification visuelle et audiovisuelle.

Organisation syllabique dans des suites de consonnes en berbère : quelles évidences phonétiques?

Rachid Ridouane & Cécile Fougeron°*
*ENST/TSI/CNRS-LTCI UMR 5141, 46, rue Barrault, 75634 Paris cedex 13
°Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins, 75005 Paris
rachid.ridouane@wanadoo.fr; cecile.fougeron@univ-paris3.fr

ABSTRACT

In this study, we examine consonants sequences in Tashlhiyt Berber in order to demonstrate that their syllabic organization can surface in their phonetic properties. Two types of three consonants sequences varying according to the degree of sonority of C1 are considered. Following syllabification principles of the language, these two types are considered to differ in their syllabic structure. Observation of the linguopalatal articulatory properties of the consonants and of the temporal coordination pattern between these consonants do show differences between the two types of sequences. These phonetic differences are interpreted as reflecting different syllabic structures, and results are confronted to the syllabification of the string proposed on phonological grounds.

1. INTRODUCTION

Il est largement admis que les segments de la chaîne parlée ne sont pas simplement juxtaposés les uns aux cotés des autres mais qu'ils sont structurellement organisés entre eux en unités plus larges. Dans ce travail nous nous intéresserons à une de ces unités : la syllabe. Une question est de déterminer s'il s'agit là d'un objet physique que l'on peut examiner et trouver les faits qui le réalisent ou s'il s'agit uniquement d'un construit théorique et abstrait utile pour les analyses linguistiques. La réalité phonétique de la syllabe, de ses composantes et des relations structurelles entre ses constituants a été débattue depuis plus d'un siècle (voir [7] pour une revue). Selon Rousselot [11 : 969] : *« La syllabe n'a rigoureusement d'existence physiologique que dans les monosyllabes isolées. Autrement, les mouvements organiques se lient les uns aux autres sans solution de continuité, et il n'y a pas de point d'arrêt dont on puisse dire d'une façon absolue : ici finit une syllabe et commence une autre.* » D'autres (par ex. Sievers, Stetson, Catford, voir [7]), au contraire, postulent que la syllabe a des corrélats physiques, bien que ceux-ci soient rarement systématiques.

Plus récemment, la Phonologie Articulatoire a contribué à ce débat en définissant la structuration syllabique des unités primitives de production (les gestes articulatoires) comme un pattern spécifique d'organisation des unités gestuelles (ex. [1, 2, 3, 4] et voir discussion). L'observation des propriétés physiques des gestes et de leur coordination permettrait ainsi d'appréhender les propriétés physiques de la syllabe et de tester expérimentalement différentes hypothèses d'organisation syllabique de la chaîne.

La langue examinée dans ce travail est le berbère chleuh. L'intérêt de cette langue réside principalement dans l'aspect typologiquement rare de sa structure syllabique, du fait de l'existence de longues suites consonantiques sans voyelles (ex. [tfktstt] *'tu l'as donnée'*, voir [10]). La question qui se pose face à de telles données est de savoir si et comment ces suites de consonnes sont organisées en syllabes. Plusieurs arguments phonologiques ont été avancés pour établir les principes qui déterminent la structure syllabique de cette langue et démontrer que des obstruentes sourdes peuvent occuper la position de noyau (voir [6, 5, 9]). Outre l'intuition des linguistes natifs, ces règles reposent sur l'application de principes universels de syllabation (relations de sonorité, préférence pour des syllabes avec attaque, contrainte contre des syllabes avec attaques ou coda branchantes, etc.), sur la prise en compte de certaines alternances morphologiquement gouvernées, et enfin sur les règles de versification.

Notre objectif, dans cette étude, est de démontrer qu'il est possible de dégager, sur des bases phonétiques, des différences dans l'organisation syllabique de diverses suites consonantiques

2. METHODE

Afin de répondre à cet objectif, nous comparons des suites de 3 consonnes $C_1C_2C_3$ formant un syntagme composé d'un préfixe (C_1) et d'une racine verbale C_2C_3V (ex : /t-kti/ '*elle se rappelle*'). Chaque racine verbale comporte deux syllabes au niveau sous-jacent /C_2.C_3V/. Deux préfixes (/t/ = 3[ème] pers. fem. singulier et /n/ = 1[ère] pers. pluriel) sont utilisés de façon à faire varier la sonorité de la consonne C_1, donnant ainsi lieu à deux conditions : la condition A où la racine verbale est préfixée avec l'occlusive sourde /t/, et la condition B préfixée de la consonne nasale /n/. En application des principes mentionnés ci-dessus, Dell & Elmedlaoui [6] proposent pour ces suites les syllabations suivantes (N=noyau, A=Attaque, C=Coda) :

(A) /$\mathbf{t}_{(A)}$ $\mathbf{k}_{(N)}$. $\mathbf{t}_{(A)}$ $\mathbf{i}_{(N)}$/ (B) /$\mathbf{n}_{(N)}$ $\mathbf{k}_{(C)}$. $\mathbf{t}_{(A)}$ $\mathbf{i}_{(N)}$/

Nous allons tester sur des bases phonétiques si ces suites de consonnes présentent une organisation syllabique différente dans les deux conditions. Nous reviendrons sur les arguments phonologiques favorisant l'une ou l'autre syllabation dans la discussion des résultats.

Six racines verbales comportant une occlusive vélaire en position C_2 et une alvéolaire en position C_3 ont été examinées dans les deux conditions (voir la table 1). Le nombre restreint de racines verbales sélectionnées répond à la contrainte d'obtenir une suite consonantique « C_1 alvéolaire + C_2 vélaire + C_3 alvéolaire» ; ceci afin de pouvoir

observer le chevauchement articulatoire entre des consonnes ayant des lieux d'articulation maximalement différents et observables sur les tracés palatographiques. Les propriétés acoustiques et l'articulation linguopalatale de ces séquences ont été observées à partir d'enregistrements électro-palatographiques (EPG 3) de la production d'un locuteur (le 1er auteur) pour 12 répétitions de chaque forme produite dans une phrase cadre : « inna jas … jat twalt » (*il lui a dit ... une fois*).

Table 1 : Matériel linguistique et conditions. Le point indique les frontières syllabiques.

Verbe		A : C1 /t/	B : C1 /n/
/k.ti/	*'se rappeler'*	tk.ti	nk.ti
/k.sa/	*'paître'*	tk.sa	nk.sa
/k.nu/	*'se courber'*	tk.nu	nk.nu
/g.za/	*'dégoûter'*	tg.za	ng.za
/g.nu/	*'coudre'*	tg.nu	ng.nu
/g.d≥i/	*'couler'*	tg.d≥i	ng.d≥i

3. HYPOTHESES, ANALYSES ET RESULTATS

Deux aspects pouvant être liés à une différence d'organisation structurelle entre les consonnes des conditions A et B sont examinés :

(a) les propriétés acoustiques et articulatoires des consonnes individuelles. Puisque C1 n'est pas comparable en conditions A et B (/t/ vs. /n/), la comparaison se limitera à C2.

(b) les propriétés de coordination temporelle entre C1 et C2 d'une part, et entre C2 et C3 d'autre part.

La comparaison statistique entre les deux conditions se fera sur (i) les valeurs obtenues pour chaque mesure, et (ii) la variabilité de chaque mesure à travers les 12 répétitions, interprétée comme un indice de stabilité. Cet indice répond à l'hypothèse selon laquelle la stabilité articulatoire ou la stabilité de la coordination temporelle entre gestes articulatoires peut témoigner de relations structurelles fortes et de liens étroits entre les éléments d'une syllabe (voir [2] et discussion).

Il est essentiel de souligner ici que l'EPG donne une mesure du contact de la langue avec le palais, et non du mouvement de la langue. Pour autant, nous analyserons ici l'évolution des profils de contacts linguopalataux à travers le temps comme des 'gestes' articulatoires en définissant certains évènements électropalato-graphiques ou acoustiques comme des pseudo-évènements articulatoires, sachant qu'ils ne sont que les conséquences de ces derniers.

3.1. Propriétés phonétiques de C2

Propriétés temporelles de C2 : Deux paramètres temporels ont été examinés : la durée acoustique de la tenue de C2 et la durée de l'occlusion linguopalatale observable sur les tracés EPG entre le début de l'occlusion complète et son relâchement. Les résultats ne montrent aucune différence significative dans les durées acoustiques et articulatoires de C2 entre les deux conditions. La variabilité de ces durées est également similaire entre les deux conditions.

Propriétés spatiales de C2 : Dans de nombreuses langues, il a été montré qu'en fonction de leur position dans la syllabe, les consonnes peuvent présenter des allophones différents ou subir des processus de lénition ou fortition affectant leur articulation. Pour examiner cela, nous avons relevé, dans un premier temps, les occurrences de C2 avec ou sans occlusion vélaire complète sur le profile EPG. Une occlusion incomplète pouvant être le reflet d'une articulation postérieure aux limites du palais artificiel ou d'une lénition de l'occlusive. Ensuite, pour les cas présentant une occlusion complète, nous avons comparé le degré de contact linguopalatal de C2 selon les deux conditions. Les résultats montrent qu'en condition A, les consonnes présentent plus fréquemment une occlusion vélaire complète. Ceci est d'autant plus vrai pour les vélaires sourdes comme le montre la figure 1. Par contre, il n'y a pas de différence de degré de contact linguopalatal, ni de différences dans la stabilité/variabilité entre les deux conditions.

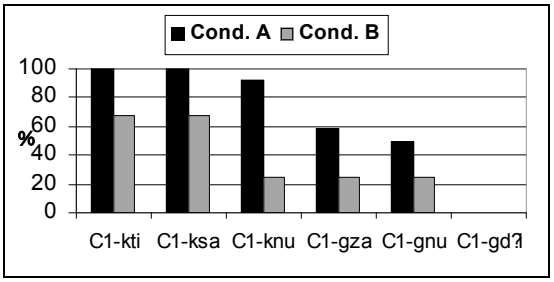


Figure 1 : Le pourcentage des cas avec occlusion vélaire complète pour C2 dans les deux conditions A et B.

Propriétés dynamiques de C2 : Dans le cadre de la Phonologie Articulatoire, les gestes vocaliques et consonantiques sont définis comme ayant des propriétés dynamiques différentes. Ainsi les gestes correspondant à des voyelles sont spécifiés par une raideur moindre comparés aux gestes consonantiques, car les articulateurs mettent plus de temps pour atteindre leur cible. Nous avons observé ici différents paramètres pouvant être considérés comme des indices de la dynamique de l'évolution du contact linguopalatal dans le temps. Notre hypothèse est que les consonnes noyaux peuvent présenter des caractéristiques propres aux gestes vocaliques. Une mesure de 'raideur' est définie comme l'intervalle temporel entre l'apparition de contact dans la région vélaire et le moment où le contact vélaire maximal est atteint (la cible). Une mesure de 'vélocité' est déterminée comme la pente de ce profil de contact (déplacement/temps). Les résultats montrent que les profils de C2 sont moins raides et de vélocité plus faible en condition B, comparés à la condition A. Aussi, les résultats montrent que le degré de contact maximal est équivalent dans les deux conditions, indiquant ainsi que ce maximum de contact est atteint moins rapidement en condition B.

3.2. Coordination temporelle entre les consonnes

Deux aspects liés à l'organisation temporelle des gestes

correspondant aux consonnes C_1, C_2 et C_3 ont été mesurés : (i) l'alignement temporel (la latence) entre différents évènements acoustiques ou electro-palatographiques définis pour les consonnes adjacentes et (ii) le degré de chevauchement entre les consonnes vélaires et alvéolaires (i.e. les intervalles temporels présentant du contact dans les deux régions). Pour se faire, différents évènements ont été définis, par ex. : les débuts d'apparition de contact dans les zones vélaire (C_2) et alvéolaire (C_1, C_3), les débuts d'occlusion, les maxima de contact dans ces régions. De plus, différentes mesures de chevauchement ont été effectuées, par ex. : chevauchement d'une consonne par la consonne adjacente rapportée à la durée totale de la 1ère consonne, à la durée de l'occlusion de cette consonne ou à la durée totale de la suite de consonnes.

Nous ne détaillerons pas ici les résultats de toutes ces mesures mais nous présenterons les tendances générales obtenues et les différences significatives.

Coordination temporelle entre C_1 et C_2 : En condition A, il y a un délai plus important entre les évènements articulatoires de C_1 et ceux de C_2. En effet, l'apparition de contacts dans la région vélaire pour C_2 est retardée par rapport à l'apparition du contact dans la région alvéolaire pour C_1. Le délai entre le début d'occlusion alvéolaire et celui de l'occlusion vélaire suivante est également plus important en condition A.

D'autre part, le chevauchement entre C_1-C_2 est plus important en condition A. Sachant que C_1 est différent dans les deux conditions (/t/ *vs.* /n/), nous avons examiné la durée des consonnes C_1, et trouvé que C_1 /t/ est plus long que C_1 /n/. Le chevauchement plus important entre C_2 et C_1/n/ qui ressort dans la condition A peut donc être lié à cette différence de durée.

La différence la plus intéressante que nous avons relevée dans ces données concerne la stabilité de la coordination temporelle entre les consonnes, mesurée par la variabilité inter-répétitions. Les différentes mesures de chevauchement ou d'alignement temporel sont moins variables en condition A qu'en condition B. Ceci apparaît par exemple si l'on compare les profils EPG des 12 répétitions des séquences /tkt/ (cond. A) et /nkt/ (cond. B) présentés sur la figure 3.

Coordination temporelle entre C_2 et C_3 : Dans les deux conditions, les consonnes C_2 et C_3 sont identiques et sont séparés au niveau sous-jacent par une frontière syllabique. Cette comparaison est donc intéressante car les différences de coordination que l'on va pouvoir observer entre les deux conditions ne devraient relever que d'une différence de position syllabique de la consonne C_2 (noyau vs. coda selon la syllabation de [6])

La comparaison de l'alignement temporel entre les différents évènements articulatoires définis pour C2 et C3 ne montre pas de différences entre les deux conditions. Par contre, il apparaît clairement que le délai entre ces différents évènements est plus variable au travers des 12 répétitions en condition B (voir par ex. sur la figure 3). L'alignement temporel entre les gestes de C2 et C3 est donc plus stable en condition A. Concernant le degré de chevauchement, il ressort de nos données qu'il y a là aussi des différences notables selon les deux conditions. En effet, moins de chevauchement entre la consonne vélaire C2 et la consonne alvéolaire C3 a été observée dans la condition A, comparée à la condition B.

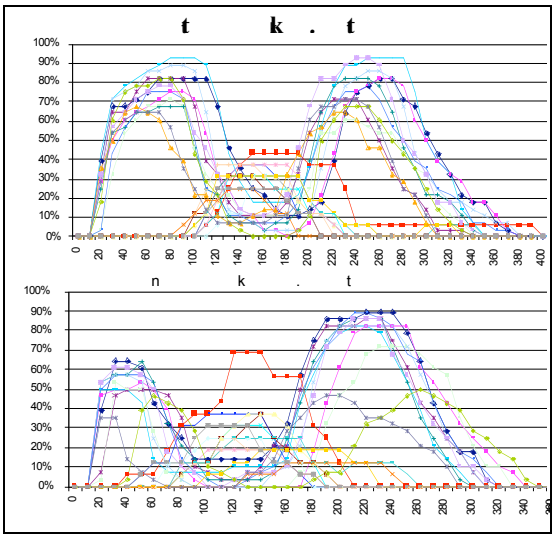


Figure 3 : Évolution temporelle du pourcentage de contact linguopalatal dans les régions alvéolaire-vélaire-alvéolaire correspondant aux consonnes C1-C2-C3. Chaque tracé correspond à une des 12 répétitions des séquences /tkt/ (cond. A, haut) et /nkt/ (cond. B, bas).

4. Discussion et Conclusion

Les différents aspects examinés dans cette étude révèlent l'existence de différences phonétiques dans les suites consonantiques examinées entre les deux conditions A et B. Si ces différences relèvent de la structuration de la chaîne comme nous en faisons l'hypothèse (au moins pour certaines, voir infra), nos résultats fournissent des arguments phonétiques montrant que dans ces deux conditions les suites de consonnes $C_1C_2C_3$ n'ont pas la même structure syllabique. Sachant que toute syllabe contient obligatoirement un noyau, et partant de l'hypothèse que les suites C_1C_2 examinées sont des syllabes à part entière, la question qui se pose est de voir si nos analyses fournissent des arguments phonétiques permettant de déterminer la structure de cette syllabe.

Revenons aux deux syllabations proposées par Dell & Elmedlaoui [6] : /$C1_{(A)}$ $C2_{(N)}$. $C3_{(A)}$ $V_{(N)}$/ en A et /$C1_{(N)}$ $C2_{(C)}$. $C3_{(A)}$ $V_{(N)}$/ en B. Ces structures syllabiques sont obtenues par l'application des principes suivants :

- en condition B, C1 /n/ étant l'élément de la suite le plus sonore, il est le meilleur candidat dans la compétition à occuper le noyau. C2 occuperait donc la position coda.
- en condition A, le principe de sonorité peut aussi s'appliquer aux séquences C_1C_2 /tg/ donnant une structure /$C1_{(A)}$ $C2_{(N)}$/. Dans les cas où C_1C_2 sont de même sonorité (/tk/), c'est l'application de principes tirés de la versification,

montrant qu'un noyau obstruant ne peut être suivi d'une coda de sonorité égale, qui favorisent une structure de type /$C1_{(A)}$ $C2_{(N)}$/.

Pouvons nous trouver dans nos données des arguments validant ces syllabations (C2 = noyau dans la condition A et C2 = coda dans la condition B) ? Il n'existe malheureusement dans la littérature que peu d'indices indiquant que les consonnes syllabiques (noyau) présentent des caractéristiques acoustiques et articulatoires permettant de les distinguer de leurs contreparties non-syllabiques. Les études sur l'anglais montrent par exemple des résultats divergents ; certains postulant que des /n/ ou /l/ syllabiques sont plus longs (ex. [8]), d'autres ne retrouvant pas de différences (ex. [12]). Concernant les données du berbère chleuh, nos comparaisons de durée ou de degré de contact ne montrent aucune différence entre nos deux conditions. Si la consonne C2 est noyau en condition A, elle n'est pas plus longue, ni articulée avec plus de contact. Par contre, la consonne C2 en condition B présente davantage de réalisations sans occlusion complète. Si l'absence d'occlusion complète sur le pseudopalais est interprétée comme de la lénition, alors C2 dans cette condition partage effectivement des caractéristiques communes aux codas dans les langues (la tendance à la lénition des codas étant fréquente).

Une autre différence est observée lorsque l'on considère les propriétés pseudo-dynamiques de C2. Si l'on considère qu'en position noyau une consonne modifie des aspects dynamiques de son articulation pour adopter des caractéristiques propres aux voyelles, alors nos résultats sous-tendraient que c'est dans cette même condition B que C2 est noyau (avec une raideur et vélocité moindre). Nous ne nous avancerons pourtant pas plus loin dans cette interprétation car les mesures considérées ne sont qu'une approximation assez lointaine de la dynamique effective du geste de la langue. Aussi, et surtout, les différences observées peuvent être fortement liées au fait que C1 est un /n/ dans la condition B et un /t/ dans la condition A.

C'est la comparaison des propriétés de coordination temporelle entre C2 et les consonnes adjacentes (C1 et C3) qui révèle à notre sens les différences les plus intéressantes. En condition A, C2 est moins chevauché par la consonne suivante et son alignement temporel est retardé par rapport à C1. Plus intéressant encore, la coordination temporelle de C2 avec les consonnes adjacentes est plus stable dans la condition A. Ainsi que ce soit en terme de chevauchement, d'alignement et de stabilité de la coordination entre les unités, la consonne C2 de la condition A semble être 'préservée' de l'influence des consonnes adjacentes. Si l'on pose l'hypothèse que dans une syllabe, l'élément 'saillant' doit être le noyau, alors cet argument supporte la syllabation /$C1_{(A)}$ $C2_{(N)}$/ en condition A.

Plusieurs travaux dans le cadre de la Phonologie Articulatoire suggèrent que les relations structurelles entre les unités se traduisent dans leur coordination et leur couplage. La stabilité de cette coordination reflèterait la cohésion entre les unités constitutives de la syllabe. Par exemple, Byrd [4] a montré que la coordination entre deux consonnes est plus stable lorsque ces consonnes occupent la position d'attaque que lorsqu'elles sont hétérosyllabiques. Cette stabilité refléterait la force du lien entre ces unités (i.e. la rigidité dans le couplage entre les gestes, voir [2]]). Il n'existe pas dans la littérature de données permettant de comparer la coordination et la cohésion des gestes dans des syllabes Attaque+Noyau *vs.* Noyau+Coda, les syllabes généralement étudiées ayant un noyau vocalique. Les suites consonantiques du berbère offrent cette possibilité. Il s'agira dans des travaux ultérieurs, basés sur les productions de plus de locuteurs, de tester plus en avant la cohésion gestuelle entre les différents composants de la syllabe.

BIBLIOGRAPHIE

[1] C. P. Browman and L. Goldstein. Gestural syllable position effects in American English. In *Producing Speech: Contemporary Issues*, F. Bell-Bertif and L. Raphael (eds.), New-York, American Institute of Physics, pp 19-33, 1995.

[2] C. P. Browman, L. Goldstein. Competing constraints on intergestural coordination and self-organization of phonological structures. *Les Cahiers de* l'ICP, *Bulletin de la Communication Parlée*, volume 5, pp 25-34, 2000.

[3] D. Byrd. C-Center revisited. *Phonetica*, volume 52, pages 285-306, 1995.

[4] D. Byrd. Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, volume 24, pp 209-244, 1996.

[5] G. N. Clements. Berber syllabification: derivations or constraints? In *Derivations and constraints in phonology,* Iggy Roca (eds.), Clarendon Press, Oxford, pp 289-330, 1997

[6] F. Dell, M. Elmedlaoui. *Syllables in Tashlhiyt Berber and in Moroccan Arabic*. Kluwer, Academic Publications, 2002.

[7] Y. Meynadier. La syllabe phonétique et phonologique : une introduction. *Travaux Interdisciplinaires du Lab. Parole et Langage d'Aix-en-Provence*, volume 20, pp 91-148, 2001.

[8] P. J. Price. Sonority and syllabicity: Acoustic correlates of perception. Phonetica, volume 37, pp 327-343, 1980.

[9] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Center for Cognitive Science Technical Report 2, 1993.

[10] R. Ridouane. Voiceless, vowel-less words in Tashlhiyt Berber: acoustic and fibroscopic evidence. Submitted.

[11] P. Rousselot. *Principes de phonétique expérimentale*. Welter, Paris, 1909.

[12] Z. Toft. The phonetics and phonology of some syllabic consonants in Southern British English. *ZAS Papers in Linguistics*, volume 28, pp 111-144, 2002.

Influence de la forme du palais sur la variabilité articulatoire

Jana Brunner[1], Pascal Perrier[2] et Susanne Fuchs[3]

[1]Humboldt-Universität zu Berlin, INP Grenoble et Zentrum für Allgemeine Sprachwissenschaft, Berlin
[2]Institut de la Communication Parlée, INPG & Univ. Stendhal, Grenoble
[3]Zentrum für Allgemeine Sprachwissenschaft, Berlin

ABSTRACT

As has been noted previously, speakers with coronally low "flat" palates exhibit less articulatory variability than speakers with coronally high "domeshaped" palates. This phenomenon is investigated by means of a tongue model and an EPG experiment. The results show that acoustic variability depends on the shape of the vocal tract. The same articulatory variability leads to more acoustic variability if the palate is flat than if it is domeshaped. Furthermore, speakers with domeshaped palates show more articulatory variability than speakers with flat palates. The results are explained by different control strategies by the speakers. Speakers with flat palates reduce their articulatory variability in order to keep their acoustic variability low.

1. INTRODUCTION

Ainsi que l'ont déjà noté un certain nombre de publications, il semble qu'il y ait une relation entre la forme du palais d'un locuteur et sa variabilité articulatoire. Par exemple, Perkell [9] a comparé les productions de six locuteurs ayant des palais différents pour les voyelles /i/, /ɪ/ et /ɛ/. Cinq de ces locuteurs avaient des palais en forme d'arche dans le plan coronal, alors que le palais du sixième était plutôt plat. Perkell a observé que ce dernier locuteur exploitait des ajustements articulatoires de très faible amplitude, tandis que des mouvements plus amples étaient trouvés pour les autres. De la même manière, Mooshammer et al. [7] ont décrit un locuteur qui présentait une variabilité articulatoire moins grande que celles d'autres locuteurs. Ce locuteur aussi avait un palais plus plat que les autres locuteurs étudiés.

Une explication possible pour ce phénomène pourrait être associée au contrôle des gestes de la parole par le locuteur en relation avec la perception de la parole produite. Il est en effet possible que les locuteurs ayant des palais plats soient obligés de réduire leur variabilité articulatoire parce que leur variabilité acoustique serait sinon trop élévée et l'intelligibilité ne serait alors plus assurée. Cette explication se justifie de la manière suivante : un locuteur avec un palais plat a un conduit vocal dont la coupe dans le plan coronal peut grossièrement être assimilée à un quadrilatère, dont le palais est la frontière supérieure, les dents ou les joues les cotés, et la langue la frontière inférieure. Au contraire, un locuteur avec un palais en forme d'arche, a un conduit vocal que l'on peut assimiler à un triangle où la langue est la base, et le palais les deux cotés. Si la position verticale de la langue change un peu, la surface du quadrilatère change plus que la surface du triangle. En d'autres termes, pour le même changement articulatoire, la fonction d'aire change plus pour un conduit vocal dont la section est proche d'un quadrilatère que pour celui dont la section est plutôt triangulaire. En conséquence, le son produit est susceptible de changer plus pour un locuteur avec un palais plat que pour un locuteur avec un palais en forme d'arche. Ceci signifie qu'un locuteur avec un palais en forme d'arche peut se permettre plus de variabilité articulatoire sans changer le son considérablement, tandis que le locuteur avec un palais plat doit articuler plus précisement pour qu'il puisse maintenir la qualité perceptive des sons produits.

Une autre question intéressante concerne la stratégie utilisée par les locuteurs pour réduire la variabilité. Pour le locuteur possédant un palais plat dans le plan coronal, il est nécessaire de trouver une position assez stable de la langue où elle ne peut pas remuer beaucoup. Une telle position est atteinte en particulier quand la langue a beaucoup de contacts palataux. Ainsi, on peut s'attendre à ce qu'un locuteur qui produit une grande quantité de contacts entre le palais et la langue ait moins de variabilité articulatoire qu'un locuteur qui a moins de contacts.

L'objectif de la présente étude est d'étudier ces relations entre la forme du palais, la variabilité articulatoire, la variabilité acoustique et les contacts linguo-palataux. Premièrement, nous étudierons la relation entre la variabilité articulatoire et la variabilité acoustique. Nous observerons en particulier si la variabilité acoustique diffère en fonction de la forme du conduit vocal pour une même variabilité articulatoire. Pour cela, nous avons effectué des simulations avec un modèle biomécanique de la langue couplé à un modèle acoustique harmonique du conduit vocal.

Ensuite, les relations entre la variabilité articulatoire, les contacts linguo-palataux et la forme du palais seront étudiées expérimentalement par électropalatographie pour 16 locuteurs allemands et norvégiens. Nous mesurerons la variabilité articulatoire, les contacts linguo-palataux et la forme du palais. Selon nos prédictions, il devrait y avoir une corrélation entre la variabilité articulatoire et la forme du palais, la variabilité étant plus grande si le palais est en forme d'arche dans le plan coronal que s'il est plat. De plus si le recours aux contacts linguo-palataux est une stratégie pour réduire la variabilité articulatoire, on peut attendre une corrélation négative entre la variabilité articulatoire et le nombre de contacts linguo-palataux.

2. MÉTHODE

La première partie de cette section traite des simulations, tandis que la deuxième partie concerne l'étude expérimentale sur les locuteurs.

2.1. Simulations

Nous avons utilisé la version la plus récente du modèle biomécanique de la langue (Payan & Perrier [8] et [12]).

La configuration de la langue pour les trois voyelles (/a/, /i/ et /u/) a été ainsi déterminée pour ce modèle.

Ces configurations de la langue ont été associées à cinq palais qui différaient dans le plan coronal. Cette courbure a été spécifiée par le coefficient α (du modèle de Heinz & Stevens [4], adapté par Perrier et al. [11]). Si α est grand, le palais est plat, s'il est petit, le palais est en forme d'arche. Pour la présente étude les valeurs suivantes ont été choisies : 3.0, 2.5, 2.0, 1.5 et 1.3. Avec cette association de la langue avec une forme spécifique du palais, les fonctions d'aire des conduits vocaux obtenus pour une configuration donnée de la langue étaient différentes selon la forme du palais. Pour cette raison, il a fallu jouer sur la position verticale de chacun de ces palais de façon à ce que les fonctions d'aire d'une même voyelle soient les mêmes quel que soit le palais considéré.

Ensuite, d'autres simulations ont été réalisées dans l'objectif de faire varier légèrement la position linguale et de simuler ainsi de la variabilité articulatoire. Ici encore, ces positions de la langue étaient combinées avec les cinq palais et les fonctions d'aire étaient calculées (pour des détails sur ces simulations cf. Brunner et al. [2]).

Finalement, à partir des fonctions d'aire, les formants des sons associés à ces différentes configurations articulatoires ont été calculés avec un modèle harmonique du conduit vocal (Badin & Fant [1]).

2.2. Expérience

16 locuteurs (10 allemands et 6 norvégiens) ont été enregistrés par électropalatographie (EPG 3.0, Reading system). Parallèlement, le signal acoustique de parole a été enregistré. Les locuteurs ont produit des phrases dans leur langue maternelle avec des logatomes qui contenaient les sons suivants : /s, ʃ, ɕ, j, i, e, ɪ, ɛ/. Chaque phrase était répétée trente fois.

Pour chacun des segments cibles le pourcentage des contacts linguo-palataux a été calculé à chaque instant d'échantillonnage des données électropalatographiques. Comme les sons avaient des durées différentes, ces pourcentages ont été interpolés puis rééchantillonnés sur 20 points quelle que soit la durée de la voyelle étudiée. Puis, la valeur moyenne et l'écart-type des trente répétitions ont été calculés pour chaque locuteur et pour chaque son à chacun des 20 points d'interpolation. Ensuite, les moyennes des 20 valeurs moyennes et des 20 écart-types ont été calculées. Si la langue est proche du palais, par exemple pour des voyelle hautes, il y a plus de contacts que s'elle est plûtot basse. En conséquence, dans le premier cas, un mouvement de la langue pourrait provoquer une variabilité du nombre de contact plus importante que dans le second. Pour limiter cette influence du nombre moyen des contacts sur la variabilité, la valeur moyenne de l'écart-type a été normalisée par la valeur moyenne du nombre de contacts. Ce coefficient a été considéré comme la mesure de la variabilité articulatoire.

L'importante du contact linguo-palatal latéral a été calculée comme le pourcentage des contacts dans la région latérale. Comme pour le pourcentage des contacts globaux, une valeur moyenne et un écart-type moyen pour les 30 répétitions et les 20 points d'interpolation ont été calculés.

La forme des palais a été caractérisée d'une part, par la mesure des coordonnées x, y et z des électrodes du palais d'électropalatographie. Puis la courbe décrite par la sixième rangée d'électrodes de ce palais a été approchée au sens des moindres carrés par une fonction exponentielle caractérisée par le coefficient alpha, conformément aux propositions de Perrier et al. [11] (cf. Brunner et al. [2] pour des détails sur les calculs mis en oeuvre).

3. RÉSULTATS

Dans la première partie les résultats des simulations seront présentés. Ils suggèrent effectivement que, pour la même variabilité articulatoire, il y a plus de variabilité acoustique si le palais est plat que s'il est en forme d'arche. La deuxième partie est consacrée aux résultats expérimentaux. Ils indiquent que les locuteurs avec des palais plats ont moins de variabilité articulatoire que les locuteurs avec des palais en forme d'arche. De plus, on observe une corrélation négative entre le nombre moyen de contacts linguo-palataux et la variabilité articulatoire. Cependant, ce résultat doit être traité avec précaution dans la mesure où la formulation même de la mesure de la variabilité, qui est normalisée par rapport au nombre moyen de contacts, introduit de facto un biais statistique.

3.1. Simulations

Sur la figure 1 on a représenté sous forme d'intervalles de confiance la variabilité des formants F1 et F2 de la voyelle /a/ générée pour les différents palais par les changements de positionnement de la langue.

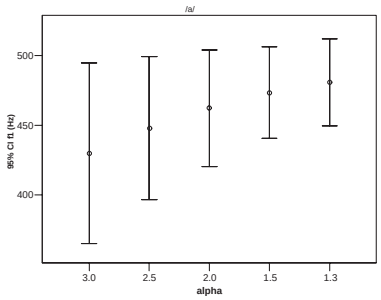


FIG. 1: Intervalles de confiance à 95% pour F1 pour les cinq palais différents. La dispersion est la plus grande pour le palais le plus plat (palais 1) et diminue pour les palais plus arqués.

C'est pour le premier palais, le plus plat avec α=3.0, que cet intervalle est le plus grand (de 365 à 494 Hz, soit 129 Hz). L'intervalle de confiance obtenu pour le palais caractérisé par α=1.3, c'est-à-dire pour le palais ayant la forme d'arche la plus marquée, est le plus petit (de 450 à 512, 62 Hz). Les autres intervalles diminuent avec α, c'est-à-dire au fur et à mesure que les palais prennent une forme d'arche plus marquée.

Les valeurs pour les autres voyelles et pour F2 sont données dans le tablaux 1. Ils vont dans le même sens que ceux de la figure 1. Les intervalles de confiance diminuent pour les palais plus arqués, c'est-à-dire ceux dont les valeurs α sont les plus petites. Il y a une seule exception, c'est le premier formant de /i/ pour lequel l'intervalle de confiance obtenu pour le palais défini par α=1.5 est plus petit que celui du palais caracterisé par α=1.3.

TAB. 1: Variabilité pour F1 et F2 (sous forme d'intervalle de confiance, en Hz) pour les cinq palais différents. La dispersion est la plus grande pour le palais le plus plat (α=3.0) et diminue pour les palais plus arqués.

	α	95% CI F1	λF1	95% CI F2	λF2
a	3.0	365-494	129	1561-1770	209
	2.5	397-499	102	1569-1757	188
	2.0	420-504	84	1583-1741	158
	1.5	440-506	66	1566-1723	157
	1.3	450-512	62	1540-1693	153
i	3.0	329-501	172	1699-2263	564
	2.5	323-492	169	1765-2283	518
	2.0	313-478	165	1842-2276	434
	1.5	304-458	145	1923-2258	335
	1.3	297-449	152	1949-2235	286
u	3.0	309-397	88	718-1553	835
	2.5	313-395	82	705-1508	803
	2.0	314-392	78	684-1453	769
	1.5	313-386	73	655-1326	671
	1.3	315-379	64	620-1272	652

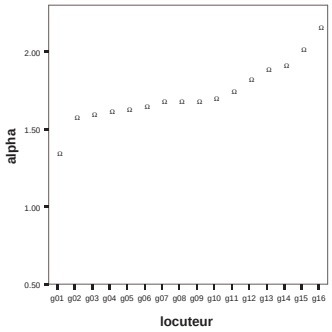


FIG. 2: Valeurs α pour tous les locuteurs. Si α est petit (à gauche) le palais est arqué dans le plan coronal, si α est grand (à droite) le palais est plat.

3.2. Résultats expérimentaux

Sur la figure 2 on peut voir les valeurs α des 16 locuteurs. Si α est petit (à gauche dans la figure) le palais du locuteur est plutôt arqué. Si, à l'opposé, α est grand, le palais est plûtot plat. Sur la figure 3 les locuteurs sont représentés dans le même ordre, mais les barres indiquent cette fois-ci la variabilité articultoire moyenne. Même si, en général, les résultats de l'expérience sont moins clairs que ceux des simulations, on peut voir qu'il y a une tendance pour les locuteurs avec des palais en forme d'arche (à gauche) à avoir plus de variabilité que les locuteurs avec des palais plats. Les barres noires indiquent les locuteurs qui ne respectent pas cette tendance générale. Les locuteurs g09 et g15 ont en effet une variabilité assez grande même si leurs palais sont plutôt plats. Les locuteurs g03 et g11, de l'autre côté, ont moins de variabilité que ne le laissait prévoir la forme de leur palais.

Une explication possible pour ces exceptions pourrait être trouvée dans leur vitesse En effet, les locuteurs g09 et g15 ont parlé avec un débit d'élocution plutôt rapide caractérisé par une durée moyenne des sons cibles étudiés respectivement égale à 77 et 82 ms, tandis que les locuteurs g03 et g11 étaient moins rapides avec une durée respective des sons cibles de 177 et 145 ms. Or d'après la loi de Fitt, bien connue dans le domaine du contrôle moteur, des gestes plus rapides sont intrinsèquement associés à une moins grande précision gestuelle, et vice-versa. On peut donc faire l'hypothèse que les contraintes induites par le débit d'élocution à chaque locuteur se superposent aux contraintes liées à la morphologie du palais jusqu'à même les contrecarrer pour ce qui concerne leurs conséquences sur la variabilité articulatoire. La comparaison de ces 4 locuteurs dont la mesure de la variabilité articulatoire va à l'opposé de nos prédictions, avec ceux parmi les autres locuteurs qui ont des palais similaires, soutient cette hypothèse. Ainsi, par exemple, les locuteurs g08 et g10, qui ont un palais aussi arqué que celui du locuteur g09, ont moins de variabilité articulatoire, et ils parlent effectivement moins vite (durée respective des sons cibles : 83 et 113 ms). Il en va de même pour le locuteur g13 comparé au locuteur g15. Ils ont des palais similaires, et présentent des durées de segments cibles respectivement égales à 118 et 74 ms. Le locuteur g13, à l'élocution plus lente, est celui des deux qui a le moins de variabilité articulatoire. Dans l'autre sens la comparaison reste aussi valable, puisque si on compare les locuteurs g03 et g11, dont la variabilité articulatoire est plus faible que ne laisse prévoir la forme de leur palais, aux locuteurs g02 et g04 (durées respectives des sons cibles : 79 et 11 ms), et g10 et g12 (durées respectives des sons cibles : 77 et 70 ms) respectivement, on constate bien que le débit d'élocution est particulièrement lent.

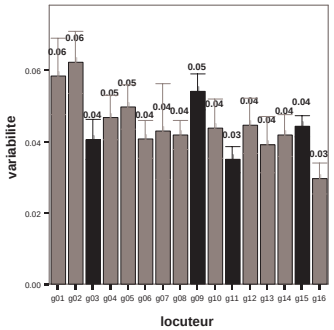


FIG. 3: Variabilité articulatoire des locuteurs présentés dans le même ordre que sur la figure 2. On observe une tendance pour les locuteurs avec des palais en forme d'arche dans le plan coronal (à gauche) à avoir plus de variabilité articulatoire que les locuteurs avec une palais plats (à droite). Les intervalles de variation correspondent à l'erreur standard.

En ce qui concerne la relation entre la variabilité articulatoire et le nombre de contacts linguo-palataux une corrélation négative a été trouvée (r=-.691, p=.000). Cependant comme nous l'avons signalé plus haut cette corrélation possède un biais statistique intrinsèque qui incite à la considérer avec précaution. En effet, pour que la mesure de la variabilité articulatoire prenne en compte le fait que naturellement la variation absolue du nombre de contacts croît avec le nombre moyen de contacts, nous avons été

amenés à normaliser cette variabilité par le nombre moyen de contacts. On corrèle donc une variable x (le nombre moyen de contacts) avec une variable k/x (la variabilité normalisée) ce qui induit de facto un résultat négatif. Effectivement, un ensemble de tests effectués sous MATLAB sur des distributions purement aléatoires du nombre de contacts et de la variabilité normalisée, nous ont montré qu'un coefficient de -0.691 est susceptible d'être expliqué par ce seul biais statistique.

4. DISCUSSION

L'objectif de la présente étude était de vérifier le phénomène selon lequel des locuteurs avec des palais plats dans le plan coronal ont moins de variabilité articulatoire que d'autres locuteurs avec des palais plus arqués, et de trouver une explication pour ce phénomène.

Les résultats des simulations indiquent que la variabilité acoustique dépend de la forme du conduit vocal telle qu'elle est déterminée par la forme du palais dans le plan coronal. Si la forme du conduit vocal dans le plan coronal est triangulaire, un changement de la position de la langue n'a pas des conséquences aussi importantes pour l'acoustique que si le conduit vocal est en forme de quadrilatère.

Les résultats de l'étude expérimentale basée sur 16 locuteurs ont démontré que les locuteurs avec des palais plats ont, en général, moins de variabilité articulatoire que les locuteurs avec des palais en forme d'arche.

En associant ces deux résultats, celui des simulations et celui de l'étude expérimentale, on pourrait suggérer que, même si la variabilité articulatoire est différente selon que les locuteurs ont des palais plats dans le plan coronal ou des palais arqués, la variabilité acoustique pourrait rester la même. En effet, nous avons montré par nos simulations avec le modèle de langue que la variabilité articulatoire des locuteurs avec des palais arqués n'a pas des conséquences aussi importantes pour la variabilité acoustique que celle des autres locuteurs. Pour les deux catégories de locuteurs, il se pourrait donc que la variabilité articulatoire soit adaptée par rapport à une contrainte de variabilité maximale dans le domaine acoustique et donc dans le domaine perceptif.

Ainsi, nos résultats, tant expérimentaux que de modélisation, soutiennent l'hypothèse que le contrôle de la précision du geste articulatoire pourrait être spécifiquement adapté à la morphologie intrinsèque du conduit vocal de chaque locuteur, de façon à préserver l'objectif ultime de la production de la parole, c'est-à-dire la qualité perceptive du son produit. Dans le célèbre débat sur la caractérisation plutôt acoustique ou plutôt gestuelle de l'espace de la tâche de la production de la parole (cf. par exemple Liberman et al. [5], Stevens [13], Lindblom [6], Fowler [3]), nos résultats plaident donc en faveur de la primauté de la composante acoustique sur la composante gestuelle (cf. Perrier [10] pour un tutorial sur cette question).

Remerciements : Cette étude est menée dans le cadre du projet POPAART financé par le CNRS et le Ministère des Affaires Etrangères français, et par la Deutsche Forschungsgemeinschaft (projet PO 334/4-1). Merci à O.Panzyga de l'université Humboldt et à A.Busler du ZAS Berlin pour la segmentation acoustique et les mesures des palais, aux locuteurs allemands du ZAS à Berlin, de l'Institut für Phonetik und Sprachliche Kommunikation de la LMU à Munich, et de l'IPDS de la CAU Kiel, et aux locuteurs norvégiens de l'Institutt for lingvistiske og nordiske studier à l'université d'Oslo, et tout spécialement à I.Moen. Pour l'aide technique merci à J.Dreyer et D.Pape.

RÉFÉRENCES

[1] Pierre Badin and Gunnar Fant. Notes on vocal tract computation. *STL-QPSR*, 2–3 :53–108, 1984.

[2] Jana Brunner, Susanne Fuchs, and Pascal Perrier. The influence of the palate shape on articulatory token-to-token variability. *ZAS Papers in Linguistics*, 42 :43–66, 2005.

[3] Carol A. Fowler. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3) :1730–1741, 1996.

[4] John M. Heinz and Kenneth N. Stevens. On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proceedings of the Fifth International Congress of Acoustic*, page A44. Liège, 1965.

[5] Alvin M. Liberman, Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74 :431–461, 1967.

[6] Björn Lindblom. Phonetic invariance and the adaptive nature of speech. In Ben A.G. Elsendoom and Herman Bouma, editors, *Working Models of Human Perception*, pages 139–173. Academic Press, London, 1988.

[7] Christine Mooshammer, Pascal Perrier, Christian Geng, and Daniel Pape. An emma and epg study on token-to-token variability. *AIPUK*, 36 :47–63, 2004.

[8] Yohan Payan and Pascal Perrier. Synthesis of v-v sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Communication*, 22 :185–205, 1997.

[9] Joseph S. Perkell. Articulatory processes. In William J. Hardcastle and John Laver, editors, *Handbook of Phonetic Sciences*, pages 333–370. Blackwell Handbooks in Linguistics, Oxford and Cambridge, Massachusetts, 1997.

[10] Pascal Perrier. Control and representations in speech production. *ZAS Papers in Linguistics*, 40 :190–132, 2005.

[11] Pascal Perrier, Louis J. Boë, and Rudolph Sock. Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract cast : modelling the transition with two sets of coefficients. *Journal of Speech and Hearing Research*, 35 :53–67, 1992.

[12] Pascal Perrier, Yohan Payan, Majid Zandipour, and Joseph Perkell. Influences that shape tongue biomechanics on speech movements during the production of velar stop consonants : A modeling study. *Journal of the Acoustical Society of America*, 114(3) :1582–1599, 2003.

[13] Kenneth N. Stevens. The quantal nature of speech : evidence from articulatory-acoustic data. In Edward E. David and Peter B. Denes, editors, *Human Communication : A Unified View*, pages 51–66. McGraw-Hill, New York, 1972.

Peut-on parler sous l'eau avec un embout de détendeur ? Etude articulatoire et perceptive

Alain Ghio, Yohann Meynadier, Bernard Teston, Julie Locco, Sandrine Clairet, Robert Espesser, Christine Meunier, Isabelle Marlien, Cyril Deniaud

Laboratoire Parole et Langage, UMR 6057, CNRS – Université de Provence,
29 Av. R. Schuman, 13621 Aix-en-Provence, France
alain.ghio@lpl.univ-aix.fr, http://www.lpl.univ-aix.fr

ABSTRACT

We study the ability of sub aquatic divers to communicate by speech by means of an air regulator mouthpiece equipped with an acoustical sensor. These specific constraints on elocution led us to carry out an aerodynamic study to check phonation, an EPG study to observe the modification of articulation, and an analysis of labial forces involved with a special mouthpiece. Tests on intelligibility enabled us to evaluate the device in situation of real diving. In the current state, the various results let foresee a reduced but real possibility of spoken communication with a mouthpiece to certain conditions.

1. LA COMMUNICATION PARLÉE SOUS-MARINE

La communication parlée sous-marine est une problématique abordée à la fois par les acteurs du domaine mais aussi par la communauté scientifique. Des workshops sur ce thème sont organisés régulièrement (ex : "European Conference on Underwater Acoustics"). De façon simplifiée, il existe actuellement trois modes de communication parlée sous-marine. (1) En caisson : les plongeurs évoluent dans une enceinte sous-pression et respirent un mélange gazeux allégé (ex : Héliox, Trimix). Les locuteurs ne sont donc soumis à aucune contrainte articulatoire. Les prises de son sont effectuées de façon traditionnelle par microphone. Par contre, la pression importante et la particularité du gaz inhalé entraîne une distorsion acoustique bien connue sous le nom d'effet "Donald Duck" [1]. (2) Avec masque facial : les plongeurs portent un masque facial leur permettant de respirer, de voir leur environnement et éventuellement de parler. Ces masques équipent notamment les plongeurs professionnels et les militaires. (3) Avec embout de détendeur : l'objet de cette étude.

2. UN EMBOUT DE DÉTENDEUR POUR PARLER SOUS L'EAU ?

Parler avec un embout de détendeur est un procédé en cours de développement qui s'adresse plutôt aux plongeurs amateurs. L'objectif est d'équiper un tel embout avec des capteurs vibratoires de façon à pouvoir enregistrer les vibrations sonores produites par le plongeur quand il parle. L'embout, en caoutchouc synthétique, fait l'interface entre le détendeur d'air comprimé et la bouche du plongeur lui permettant de respirer. Les enjeux industriels nous imposent une réserve de confidentialité et nous ne rentrerons pas dans les détails du capteur de vibration employé. Il faut toutefois signaler que ce capteur et son conditionneur entraînent certaines distorsions, notamment un filtrage passe-haut avec une perte de -15 dB à 250Hz pour une référence 0dB à 1000Hz. On peut donc s'attendre à certains problèmes de transmission pour les voix graves. Notre objectif était de réaliser une étude articulatoire et perceptive sur différents types d'embout et d'en évaluer les impacts sur la production et la perception de la parole en plongée.

L'embout standard (*EmbStd*) est la version classiquement utilisée en plongée. Pour la production de parole, sa forme entraîne une perturbation des articulations antérieures, notamment une impossibilité d'occlusion labiale et une réduction de mouvement de la mâchoire. Une variante d'embout utilise une forme standard à laquelle se rajoute une "pièce phonatoire" en matière plastique (*EmbPPhon*) qui permet un pincement du tube au niveau des dents. Ce système "prothétique" a été proposé pour autoriser une occlusion "bi-dentale" de type "morsure". Enfin, une forme originale a été testée intégrant un système de clapet au niveau des lèvres autorisant, en théorie, une pseudo fermeture labiale (*EmbClapet*).

3. LES CONTRAINTES PHYSIOLOGIQUES LIEES A LA PRODUCTION DE PAROLE AVEC EMBOUT

Cette situation atypique de production de parole entraîne des contraintes à la fois aérodynamiques (air comprimé) et articulatoires (blocage des lèvres et de la mâchoire).

3.1. La surpression aérienne

Position du problème

L'utilisation de l'ensemble embout + détendeur + air comprimé perturbe inévitablement les processus naturels de phonation, notamment en introduisant une surpression en sortie du conduit vocal. L'objectif est d'évaluer l'ampleur de la perturbation et de vérifier si cette surcharge n'est pas prohibitive à un usage de la parole en milieu subaquatique. Un paramètre essentiel pour la phonation est la pression transglottique qui permet une mise en vibration des cordes vocales. Une élévation de la pression sus-glottique (liée à l'air comprimé) doit nécessairement être accompagnée d'une augmentation de la pression sous-glottique (PSG), ceci pour maintenir le différentiel de pression nécessaire à la vibration laryngée. La mesure de la PSG est donc essentielle pour évaluer la faisabilité du mécanisme. En milieu clinique, il est possible de ponctionner la trachée pour faire passer un cathéter pour la mesure. Dans notre cas (milieu aquatique), cette méthode invasive est impossible. De ce fait, nous avons utilisé la méthode indirecte de "Smitheran & Hixon" validée par Demolin et al. [2]. Par ce procédé, on accède à la pression sous-glottique en mesurant la pression intra-orale lors de configurations articulatoires bien spécifiques (ex : la tenue de /p/).

Mesures, résultats et conclusion

Trois plongeurs ont prononcé la phrase "Pa pa ne m'a pas parlé de beau-pa pa" dix fois, soit 60 productions de /pa/ par locuteur, ceci une fois à l'air libre sans embout (condition de contrôle) puis en milieu aquatique avec l'embout. La pression intra-orale a été mesurée à l'aide d'une sonde d'aspiration insérée dans la bouche du locuteur et reliée au capteur de pression du dispositif EVA2 (SQLab, LPL). L'utilisation de l'ensemble embout + détendeur + air comprimé laisse apparaître (Figure 1) un surcroît de pression sous-glottique de l'ordre de 4 hPa (contrôle : 5.13 hPa, aquatique : 9.03 hPa). Cette valeur est importante car elle représente une augmentation de 80%. Toutefois, la charge supplémentaire induite sur la source vocale n'est pas prohibitive et autorise la phonation. En effet, dans le cas de parole pathologique, la pression sous-glottique estimée peut atteindre des valeurs de l'ordre de 20 hPa.

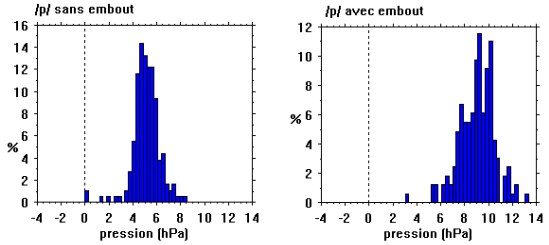


Figure 1 : Histogramme des mesures de pression tous locuteurs confondus en condition normale (sans embout à l'air libre) et avec embout sous l'eau.

3.2. La mesure de force labiale

En utilisant l'embout doté d'un clapet à ressort permettant, par construction, une pseudo fermeture labiale, les utilisateurs ont très vite eu la sensation que les forces à développer au niveau des lèvres étaient trop importantes pour assurer une occlusion, même partielle. Afin de vérifier cette sensation, nous avons développé un banc de mesure pour obtenir une valeur des forces mises en jeu. Pour cela, nous avons utilisé des capteurs de force Interlink Electronic (réf. SS-U-N-S00033) dont la résistance électrique varie en fonction de la force qui lui est appliquée (FSR). Nous avons choisi ce type de capteur car il est insensible aux vibrations et donc au bruit environnant contrairement aux polymères piézo-électriques; de plus, sa large plage d'impédance autorise l'emploi d'une électronique d'interface simplifiée. La mesure de résistance à partir d'un ohmmètre de qualité permet ainsi de déterminer la force. Une étape d'étalonnage a été effectuée au préalable pour obtenir la relation résistance/force.

Table 1 : Force labiale pour obtenir une occlusion

Condition	Force
Sans embout	0.4 N
Avec embout à clapet	4.9 N

Les résultats (Table 1) laissent apparaître des écarts de force d'un rapport 10 entre la situation naturelle et celle avec l'embout à clapet, ce qui objective l'importante rigidité perçue et qui implique un développement de force beaucoup trop important pour une utilisation opérationnelle du clapet.

3.3. Etude articulatoire par EPG

L'effet d'embouts, maintenus au niveau prémolaire, sur l'articulation linguopalatale n'est pas complètement assimilable à celui de bite-blocks (cales intermolaires). En effet, un embout de plongée est plus invasif dans la partie antérieure de la bouche que des bite-blocks, mais laisse une certaine liberté de mouvement mandibulaire. On peut cependant s'attendre à observer des phénomènes proches de ceux isolés par Clairet [3] sur l'effet bite-block en français, à savoir une postériorisation du lieu articulatoire de /t/ sans incidence notable sur le mode occlusif.

Méthode et mesures

L'électropalatographie (EPG) enregistre dans le temps les contacts de la langue sur le palais dur dans les dimensions sagittale et coronale grâce au port d'un palais artificiel garni d'électrodes qui s'activent au contact de la langue (dans notre cas, EPG de Reading à 62 électrodes, [4]). Le corpus est constitué de séquences CaCa, avec alternativement /p, t, k, f, s, ʃ, j, w/, insérées dans la phrase porteuse « Il reverra 'CaCa' à Draguignan » répétée 3 fois par 3 locuteurs français. 3 indices EPG sont calculés au point de constriction maximale de chaque consonne. Le *Taux Maximal* de contacts EPG, croissant de 0 à 1, renseigne la magnitude de la constriction linguopalatale. L'*Indice de Postériorité*, qui varie de -3,5 à +3,5 plus le contact linguopalatal est postérieur, informe sur le lieu d'articulation. L'*Indice de Fermeture*, intégrant une pondération croissante de 1 à 3 des 2 colonnes d'électrodes les plus latérales aux 2 colonnes les plus centrales, augmente plus le mode d'articulation est fermé. Pour ces calculs, les 2 colonnes EPG les plus latérales ont été exclues du fait d'un recouvrement partiel et variable par les différents embouts.

Résultats

L'effet général d'un embout sur l'articulation linguopalatale est fonction de la forme de l'embout et des locuteurs (Figure 2). Une ANOVA à 2 facteurs (*Locuteur*Embout*) montre que les locuteurs [$F(2, 432) = 8.307$; $p = .003$] et les embouts [$F(3, 432) = 2.738$; $p = .0431$] ont un effet significatif, mais variable, sur la magnitude de la constriction linguopalatale.

L'Indice de Postériorité, relatif à la localisation des contacts EPG sur l'axe antéro-postérieur du palais, révèle que le port d'un embout provoque globalement une postériorisation du lieu d'articulation des consonnes linguopalatales (Figure 3). Face à ce phénomène, les antérieures, [t] dental et [s] alvéolaire, sont les plus affectées. Les palatales [ʃ, j] et les vélaires [k, w] montrent une postériorisation moins importante et plus variable en fonction du type d'embout. Des tests post-hoc Bonferroni/Dunn (seuil calculé à $p < .0083$) réalisés entre les différents conditions d'embout (2 facteurs intégrés, *Lieu*Embout*) montrent que les seules distinctions significatives concernent la condition *Contrôle* face à toutes celles avec embout ($.0001 < p < .0006$).

L'Indice de Fermeture (Figure 4), relatif au degré de fermeture de la constriction linguopalatale (axe sagittal),

montre que l'embout n'entraîne pas de fusion de mode d'articulation entre les approximantes [j, w], les fricatives [s, ʃ] et les occlusives [t, k]. Cela provient en partie d'un effet opposé entre occlusives (nettement plus fermées) et fricatives (légèrement plus ouvertes). Des tests post-hoc Bonferroni/Dunn (seuil calculé à $p < .0083$) réalisés entre les différents conditions d'embout (2 facteurs intégrés *Mode*Embout*) montrent que la seule distinction significative observée concerne la condition *Contrôle* face à *EmbClapet* ($p = .0056$). Cela apparaît sur la Figure 4 par une quasi fusion des modes approximant et fricatif dans cette condition. Ce type d'embout semble donc provoquer une gêne plus importante.

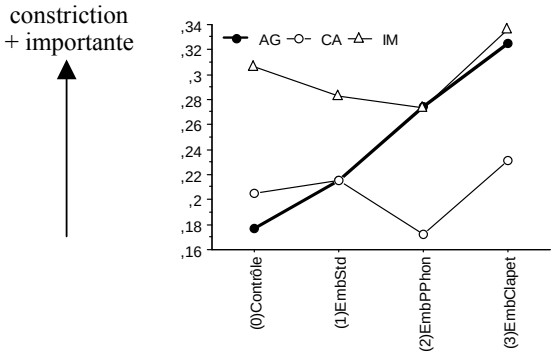


Figure 2: *Taux Maximal de contacts* moyen (en y) selon le type d'embouts et les 3 locuteurs (AG, CA, IM)

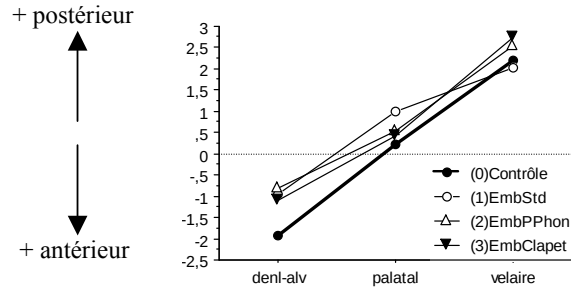


Figure 3: *Indice de Postériorité* moyen (en y) selon le lieu d'articulation et les embouts (locuteurs confondus)

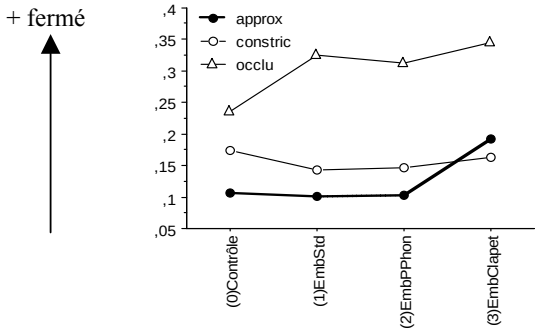


Figure 4: *Indice de Fermeture* moyen (en y) selon le mode d'articulation et les embouts (locuteurs confondus)

L'analyse EPG des articulations labiales [p, f] fait apparaître quelques contacts latéraux postérieurs dans les conditions *EmbPPhon* et *EmbClapet*. Ce phénomène récurrent semblerait se produire quand un geste d'occlusion labial substitutif est proposé par une adaptation de l'embout autorisant une fermeture de son orifice, soit au niveau des dents par une pièce phonatoire actionnée en mordant, soit par un clapet actionné en pinçant au niveau des lèvres.

Pour les consonnes linguopalatales, l'analyse EPG a donc mis en évidence des modifications articulatoires non réellement critiques. Seul l'embout à clapet entraîne une neutralisation de la distinction de lieux ou de modes articulatoires entre les consonnes. Le changement le plus homogène est lié à une postériorisation de /t, s/ ce qui peut s'apparenter à un effet bite-block [3]. Reste que l'effet *Locuteurs* demeure le plus important lors des adaptations articulatoires provoquées par la perturbation que représente le port d'un embout de plongée.

4. L'EVALUATION PERCEPTIVE (MESURE D'INTELLIGIBILITE)

L'intelligibilité du dispositif a été mesurée par des tests de paires minimales (Peckels & Rossi, [5]).

4.1. Recueil du corpus

Trois plongeurs (1 homme, 2 femmes) ont prononcé avec chacune des formes d'embout la liste des 216 mots isolés correspondant aux 108 paires minimales. Un enregistrement en chambre sourde sans embout a fourni les données de contrôle. Les enregistrements aquatiques ont été réalisés en piscine. Les plongées ont été placées sous le contrôle d'un moniteur de plongée fédéral (MF1). Le texte du corpus a été imprimé sur des transparents résistant à un trempage dans l'eau chlorée. Le signal émis par le capteur de l'embout était pré amplifié, enregistré sur Mini-Disque puis stocké au format audio Wave PCM 16 bits 44kHz.

4.2. Tests d'intelligibilité

Pour chaque embout et chaque locuteur, 10 auditeurs ont fourni leur évaluation (2160 réponses) à travers une tâche de choix forcé portant sur chaque paire minimale [5]. Les 4 conditions (1 contrôle + 3 embouts) avec trois locuteurs ont permis de recueillir 25 920 réponses. Le pilotage expérimental était assuré par le logiciel PERCEVAL, un dispositif automatisé de tests de PERCeption et d'EVALuation auditive et visuelle [6].

La situation de contrôle (98.8% de bonnes réponses) a permis de valider la "bonne" élocution des locuteurs et la perception "correcte" des auditeurs (Figure 5). Les technologies "simples" (embout standard avec/sans pièce phonatoire) ont fourni les meilleurs résultats avec 78% de réponses correctes (Figure 5). Cette tendance est confirmée par l'analyse du temps de réaction où les mêmes embouts nécessitent un temps de réponse plus court (1103 ms vs. 1324 ms avec l'embout à clapet, Figure 5). Ce taux de 78% de bonnes réponses reflète une situation mitigée (il faut se rappeler qu'un score de 50 % est équivalent à une réponse au hasard pour des paires de mots). A titre de comparaison, le score de bonnes réponses dans la communication en milieu subaquatique hyperbare se situe entre 73 et 78 % dans l'étude de Cavé et al. [1] et d'après les auteurs, un tel score ne pose pas de problèmes aux plongeurs professionnels qui considèrent qu'une telle communication est satisfaisante.

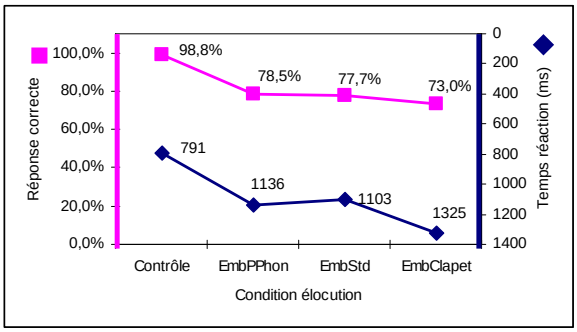


Figure 5 : Taux de réponses correctes (■) et temps de réaction (◆) par embout (locuteurs confondus)

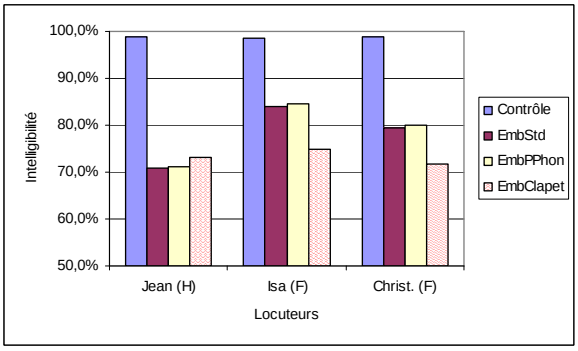


Figure 6 : Scores d'intelligibilité par locuteur selon la situation d'enregistrement

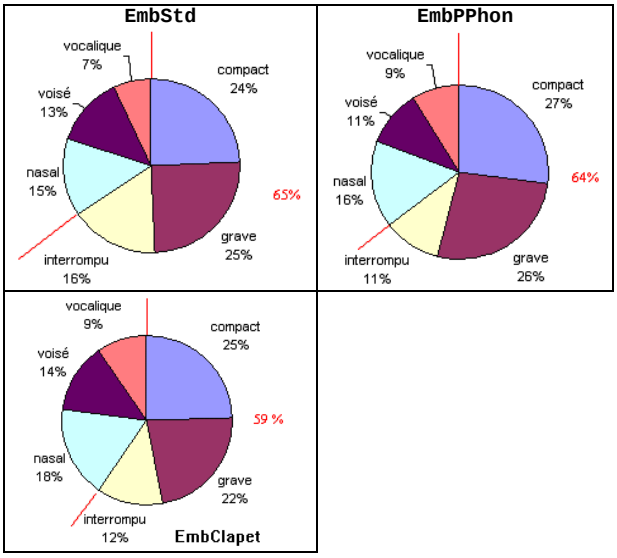


Figure 7 : Répartition des erreurs par embout (locuteurs confondus)

L'analyse des résultats par condition et par plongeur (Figure 6) laisse apparaître un effet important lié au locuteur, plus qu'à la technologie employée. Autrement dit, l'amélioration de l'intelligibilité repose, dans un premier temps, plus sur une mise en place de stratégies compensatoires et un apprentissage rapide chez le plongeur qu'à la technologie de l'embout, cette stratégie étant d'emblée disponible plus chez certains locuteurs que chez d'autres.

Quels que soient les embouts, les traits phonétiques les plus altérés sont le trait de compacité et le trait grave/aigu (Figure 7). Ces deux traits sont des corrélats acoustiques du lieu d'articulation (grave ⇔ labialité, aigu ⇔ dental, compact ⇔ palatal). Les erreurs sur le trait "interrompu" sont moins fréquentes avec la pièce phonatoire et l'embout à clapet, ce qui traduit un fonctionnement opérationnel de ces dispositifs d'occlusion.

5. CONCLUSION

Au niveau scientifique, ce type d'étude reste très difficile à mener du fait des contraintes expérimentales très lourdes et inhabituelles. Malgré tout, les résultats en perception rejoignent en bonne partie ceux de l'étude articulatoire. La postériorisation des consonnes linguales, et principalement antérieures, provoque globalement un rapprochement général des lieux, ce qui correspond auditivement à des distinctions perceptives plus difficiles pour les traits grave/aigu et compact/diffus. Les tests perceptifs montrent également relativement peu de confusion entre le mode occlusif et les autres. Cela semble corroborer par l'observation en production d'occlusives nettement plus fermées et de constrictives un peu plus ouvertes, ce qui favorise un maintien perceptif de la distinction [+/-interrompu]. Ainsi, en l'état, les différents résultats laissent entrevoir une possibilité réelle, même si elle s'avère réduite, de communication parlée subaquatique avec un embout de détendeur.

Cette étude a fait l'objet du contrat de prestation CNRS n°2003 IND 073.

Remerciements à la direction et au personnel de la piscine universitaire d'Aix-en-Provence.

BIBLIOGRAPHIE

[1] C. Cavé, C. Meunier, A. Ghio, Melliet J.L., Marchal A. Effect of speech conditions and gas mixture on the intelligibility of Diver's speech, as asserted under real Diving Conditions at 50 and 100 meters, In *3rd European Conference on Underwater Acoustics*, pages 765-769, 1996.

[2] D. Demolin, A. Giovanni, S. Hassid, C. Heim, V. Lecuit, A. Socquet., Direct and indirect measurements of subglottic pressure, In *Larynx 97*, Marseille, pages 69-72, 1997.

[3] Clairet, S., Compensation articulatoire dans la production des occlusives linguales du français, In *Colloque "Perturbation et réajustements : langue et parole"*, Haguenau, 2005.

[4] W.J. Hardcastle, W.J. Jones, C. Knight, A. Trudgeon, C. Calder, New developments in Electropalatography: State-of-the-art report, In *Clinical Linguistics and Phonetics*, 3 : 1-38, 1989.

[5] J.P. Peckels & M. Rossi. Le test de diagnostic par paires minimales, *Revue d'acoustique*, 27 : 245-262, 1973.

[6] C. André, A. Ghio, C. Cavé, B. Teston. PERCEVAL : a Computer Driven System for Experimentation on Auditory and Visual Perception. In *15th ICPhS*, pages 1421-1424, 2003.

Production des voyelles nasales en français québécois

Delvaux Véronique

FNRS, Université de Mons-Hainaut
18, Place du Parc, 7000 Mons, Belgique
delvaux@umh.ac.be
http://www.staff.umh.ac.be/Delvaux.Veronique

ABSTRACT

This paper aims at describing the production of nasal vowels by 5 speakers of Canadian French (Montreal). The data consist in images of the tongue that have been tracked by ultra-sound while simultaneously recording nasal airflow with PcQuirer and the movements of the lips using a video camera. Results show that: (i) nasalization is delayed in Canadian French nasal vowels (especially /ɛ̃/); (ii) the majority of the vowels are diphtongized, diphtongization being larger in front vowels than in back vowels and in closed syllables than in open syllables. The way Canadian French deals with the constraints acting upon nasalization are discussed, especially in comparison with European French.

1. INTRODUCTION

La nasalité vocalique implique l'interaction de nombreux facteurs phonétiques et phonologiques. En production de la parole, la coproduction du geste d'ouverture du port vélo-pharyngal (VP) avec les autres gestes dans le conduit vocal amène à une grande variabilité dans les langues du monde quant à l'amplitude et au timing des mouvements articulatoires, ce qui constitue l'un des enjeux essentiels pour les théories de la coarticulation. Par ailleurs, de nombreuses contraintes, notamment d'ordre perceptuel, agissent sur l'implémentation de la nasalité vocalique. Ainsi, les effets acoustiques du couplage nasal impliquent une réduction de l'espace perceptuel pour les voyelles nasales par rapport aux orales correspondantes [1]. En particulier, la perception de la nasalité interagit avec la sensation de hauteur pour les voyelles. Sur l'axe syntagmatique, la perception de la nasalité vocalique dépend du taux de nasalisation des consonnes voisines [2]. Elle interagit également avec le degré d'ouverture glottique des consonnes environnantes [3]. Enfin, des contraintes externes peuvent agir sur la réalisation de la nasalité, notamment celles qui sont liées à l'inventaire phonologique de la langue: les voyelles nasales doivent être suffisament différentes phonétiquement des orales qui leur correspondent phonologiquement.

Le français fournit un cas d'étude particulièrement intéressant parce que: (i) le contraste phonologique de nasalité y est présent tant pour les voyelles que pour les consonnes, ce qui provoque de nombreux phénomènes de coproduction-coarticulation dans la chaîne parlée; (ii) le système phonologique des locuteurs peut compter jusqu'à 16 voyelles /i, e, ɛ, a, ɑ, ɔ, o, u, y, ø, œ, ə, ɛ̃, ɑ̃, ɔ̃, œ̃/, ce qui peut induire de nombreuses interactions perceptuelles entre articulations covariantes.

Un intérêt supplémentaire du français est que les voyelles nasales sont réalisées phonétiquement de façon nettement différente dans les trois grands groupes dialectaux: le français québécois, le français européen septentrional (moitié nord de la France, Belgique, Suisse) et le français méridional. L'étude comparée de ces réalisations doit permettre de déterminer comment chacun des dialectes a rencontré les différentes contraintes en relation avec son système propre, et d'ainsi aborder la question du niveau de traitement cognitif de ces phénomènes: règles phonologiques spécifiques, règles phonétiques, ou bien phonétique contrôlée [4]? Les données présentées ici s'inscrivent dans ce cadre de recherche. L'objectif de l'article est de décrire la production des voyelles nasales en français québécois (Montréal) au moyen des techniques de la phonologie de laboratoire. Jusqu'ici, la plupart des travaux rapportés dans la littérature sont basés sur le jugement de phonéticiens entraînés et/ou sur des mesures exclusivement acoustiques [5], ou encore sur des données articulatoires acquises par des techniques aujourd'hui dépassées ([6] entre autres).

2. MATÉRIEL ET MÉTHODE

2.1. Sujets et corpus

Six locuteurs francophones âgés de 22 à 29 ans ont participé à l'expérience. Il s'agit de cinq locuteurs québécois (trois hommes: AC, LH, MC et deux femmes: AL, CE) et d'une locutrice francophone belge (VD). Le corpus est constitué de 42 séquences de non mots, réparties en quatre groupes (voir table 1).

Table 1 : Corpus.

Groupe 1	Groupe 2	Groupe 3	Groupe 4
ɑɑ̃ɑ	mɑ pɑ	pɑ ɑm	mɑm mɑ̃
ɛɛ̃ɛ	mɛ pɛ	pɛ ɛm	mɛm mɛ̃
ɔɔ̃ɔ	mɔ pɔ	pɔ ɔm	mɔm mɔ̃
œœ̃œ	mœ pœ	pœ œm	mœm mœ̃
	mi pi	pi im	pɑ̃ ɑ̃p
	my py	py ym	pɛ̃ ɛ̃p
	mu pu	pu um	pɔ̃ ɔ̃p
	me pe	pe em	pœ̃ œ̃p
	mø pø	pø øm	sɑ̃ ɑ̃s
	mo po	po om	sɛ̃ ɛ̃s
	ma pa	pa am	sɔ̃ ɔ̃s
	mɑ pɑ	pɑ ɑm	sœ̃ œ̃s

Chaque groupe a fait l'objet d'une acquisition séparée, au cours de laquelle les séquences ont été répétées trois fois. En ce qui concerne le groupe 1, on a demandé aux locuteurs de prononcer les séquences en une seule expiration, sans marquer d'arrêt entre les voyelles, afin de comparer la position des articulateurs entre les orales et les nasales correspondantes.

2.1. Acquisition des données

Les données ont été acquises au Laboratoire de Phonétique de l'Université du Québec à Montréal dans une pièce isolée acoustiquement. Quatre types de paramètres ont été enregistrés simultanément.

(i) La position des lèvres a été obtenue en filmant le locuteur de profil au moyen d'une caméra digitale (30 images/sec). L'image a été volontairement surexposée de façon à maximaliser le contraste entre les lèvres et le reste du visage. De plus, les lèvres étaient maquillées en bleu, ce qui permet une distinction maximale avec la peau (pigmentée en rouge). Le zoom était adapté à chaque locuteur; une grille placée devant l'objectif en début d'enregistrement a permis d'étalonner les images et de comparer ainsi les données sujet à sujet.
(ii) Des coupes médio-sagittales de la langue ont été obtenues par ultra-sons. Il s'agit d'un appareil à usage médical (Sonosite 180) dont les images sont acquises en temps réel via une sonde placée sous la mâchoire inférieure du sujet. Ces images ont été directement enregistrées au moyen d'une seconde caméra digitale (30 images/sec). La position de la tête du locuteur était fixée par le port d'un casque arrimé à la cloison murale. La sonde à ultra-sons était elle-même fixée sur un pied positionné sous le menton du sujet. Une épaisseur de gel d'environ 1 cm a été déposée sur la sonde de façon à permettre les mouvements de la mâchoire inférieure sans modification de la position de la sonde. A la fin de la session expérimentale, on a demandé au locuteur d'avaler lentement une gorgée d'eau afin de déterminer la position du palais et de la reporter sur les images correspondantes.
(iii) Le débit d'air nasal a été mesuré à la sortie du masque nasal par la station de travail PCQuirer.
(iv) Le signal de parole a été enregistré au moyen d'un microphone externe branché à la seconde caméra.

2.2. Traitement des données

Les différents signaux ainsi acquis ont été resynchronisés temporellement puis visualisés dans le logiciel '4ChannelsExplorer', une application personnalisée réalisée par l'auteur à partir du logiciel éditeur de médias iShell: www.tribeworks.com (voir figure 1). La post-synchronisation a été rendue possible par l'introduction régulière d'un signal sonore au cours de l'acquisition des données, signal qui a laissé une empreinte sur les pistes audio des deux caméras vidéo et de PCQuirer. Le corpus a été segmenté sur la base des informations fournies par la forme d'onde et un spectrogramme. Les enregistrements des deux caméras digitales ont ensuite été exportés sous la forme de séquences d'image qui ont été traitées par le logiciel EdgeTrack afin d'en extraire les contours (http://vims.cis.udel.edu/EdgeTrack).

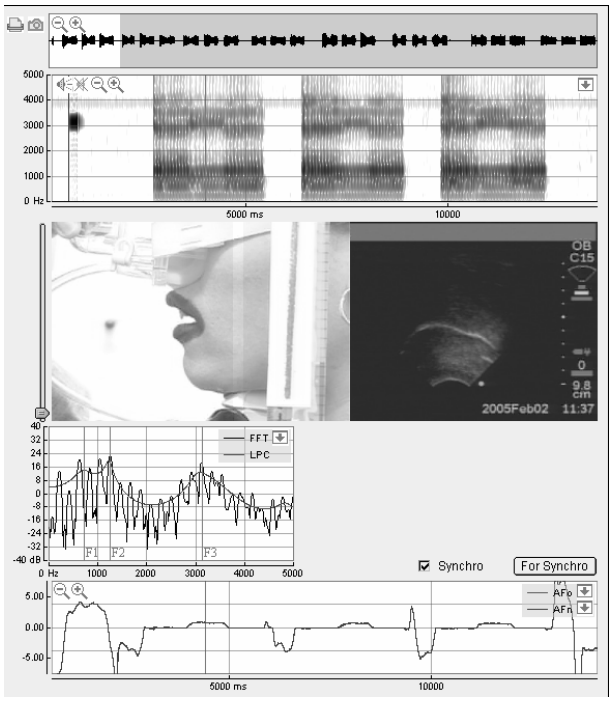


Figure 1: Visualisation des données.

3. RESULTATS

3.1. Données aérodynamiques

Par rapport aux voyelles nasales du français européen, les nasales québécoises sont moins nasalisées: une portion significative de la nasale peut être prononcée avec un débit d'air nasal (DAN) égal à zéro. L'ampleur de cette dénasalisation varie selon la voyelle concernée et le contexte phonologique, comme le montre la table 2. La table 2 donne le pourcentage de la durée de la voyelle pour laquelle le DAN est égal ou inférieur à zéro en fonction du dialecte (Québec vs. Belgique), du contexte (items N$\tilde{V}$ vs. C$\tilde{V}$ vs. $\tilde{V}$C), et de la voyelle. On voit que la dénasalisation est plus marquée pour les voyelles antérieures /œ̃/, et surtout /ɛ̃/, que pour les postérieures /ɑ̃, ɔ̃/. Par ailleurs, les contextes N$\tilde{V}$ et $\tilde{V}$C sont les plus favorables à la dénasalisation. Le phénomène est de moindre ampleur, mais toujours présent, en contexte C$\tilde{V}$. Ici les valeurs sont comparables à celles obtenues avec la locutrice belge sauf dans le cas de /ɛ̃/. Il faut noter que dans les items $\tilde{V}$C et C$\tilde{V}$, c'est le début de la voyelle qui n'est pas nasalisé car le DAN tarde à se mettre en régime. En contexte N$\tilde{V}$ par contre, la voyelle est d'abord nasalisée dans la foulée de la consonne nasale, puis le DAN redescend jusqu'à zéro, avant de remonter dans le dernier quart de la voyelle jusqu'à une valeur dépassant celle atteinte pour N. Ainsi, toutes les nasales québécoises sans exception sont marquées par une finale très nasalisée.

Table 2 : Pourcentage de la duréee de la voyelle nasale pour laquelle le DAN est inférieur ou égal à zéro.

		Québec				Belgique			
		N$\tilde{V}$	C$\tilde{V}$	$\tilde{V}$C	All	N$\tilde{V}$	C$\tilde{V}$	$\tilde{V}$C	All
/ɛ̃/	M	58	49	60	56	2	16	23	14
	SD	8	11	9	9	2	4	4	3
/œ̃/	M	47	16	55	39	0	16	18	11
	SD	10	8	13	10	3	6	3	4
/ɑ̃/	M	6	12	23	14	0	14	20	11
	SD	4	8	7	6	1	7	8	5
/ɔ̃/	M	22	6	26	14	0	12	13	13
	SD	6	4	5	5	0	2	4	2
All	M	33	21	41	31	1	14	18	12
	SD	7	8	9	8	1	5	5	6
	N	60	120	120	300	12	24	24	60

3.2. Données articulatoires et acoustiques

Nos données confirment que la caractéristique principale des voyelles nasales québécoises par rapport à leurs homologues européennes est la diphtongaison. Le phénomène est plus marqué en syllabe fermée qu'en syllabe ouverte, soit N$\tilde{V}$ < C$\tilde{V}$ < $\tilde{V}C_{[p]}$ < $\tilde{V}C_{[s]}$. Dans les contextes N$\tilde{V}$ et C$\tilde{V}$, la diphtongaison est notable pour les antérieures, mais elle est peu manifeste pour /ɔ̃/ et surtout pour /ɑ̃/. Ainsi, la figure 2 donne les résultats obtenus pour la production des quatre voyelles nasales par la locutrice québécoise CE en contexte C$\tilde{V}$, soit de haut en bas: /pɛ̃, pœ̃, pɑ̃, pɔ̃/. Les quatre types de données sont rassemblés dans la figure 2. A gauche sont affichés le tracé de débit d'air nasal ainsi qu'un spectrogramme à bande large. Trois repères temporels ont été placés sur le signal: (1) début de l'occlusive sourde (on a pris pour repère le passage à zéro du DAN après respiration) ; (2) mise en vibration des cordes vocales ; (3) fin de la vibration des cordes vocales. A droite de la figure sont superposés tous les contours de langue et de lèvres obtenus entre les repères temporels (1) et (3) (et englobant ces repères, étant entendu qu'un contour correspond à une portion de 33 ms de signal). Dans ce cas-ci, il y a en tout 14 contours pour /pɛ̃/, 15 pour /pœ̃/, 13 pour /pɑ̃/ et 16 pour /pɔ̃/ et le cinquième contour correspond au repère (2) à chaque fois. Les flèches indiquent la direction des changements observés dans la position des articulateurs au cours de la voyelle, soit entre (2) et (3). Ainsi, au niveau des lèvres, les premiers contours sont toujours différents des autres parce qu'ils dénotent la tenue puis le relâchement de l'occlusion bilabiale (entre (1) et (2)). Au cours de la voyelle proprement dite, on n'observe des modifications que dans le cas de /ɛ̃/ et /ɔ̃/. Pour /ɔ̃/, il s'agit d'un arrondissement des lèvres, surtout perceptible pour la lèvre inférieure, et qui survient simultanément à une légère postériorisation de la langue: [ɔ̃o͂]. Les deux mouvements ont pour effet d'abaisser la fréquence de F2 au cours de la nasale. Dans le cas de /ɛ̃/, la langue entame un mouvement vers le haut et vers l'avant de la bouche pendant la consonne /p/, mouvement qui se poursuit au cours de la voyelle, soit [æɛ̃ʲ]. D'ailleurs, F2 suit une trajectoire montante pendant toute la voyelle; celle-ci se termine sur une phase bruitée nasalisée où la lèvre inférieure est retroussée et on voit la langue pointer entre les lèvres sur la vidéo: [ʲ]. Pour /œ̃/, l'évolution temporelle de la langue (et de F2) est comparable à celle observée pour /ɛ̃/, soit [œ̃ø̃ʷ̃] Enfin, il y a peu de diphtongaison de /ɑ̃/ en contexte C$\tilde{V}$, alors que les items /ɑ̃s/ étaient régulièrement réalisés [aɑ̃ʷ̃s] par 4 locuteurs sur 5.

4. DISCUSSION ET CONCLUSION

La réalisation phonétique des nasales par nos locuteurs québécois peut être qualifiée d'hybride par rapport aux réalisations des deux autres dialectes. Du point de vue du couplage nasal, l'ouverture du port VP est retardée par rapport à ce qui est observé en français septentrional (cf. p.ex. [7]), mais une portion significative de la voyelle reste nasalisée et on n'observe qu'un très court appendice (semi-)consonantique nasal, moins marqué qu'en français méridional. Du point de vue des autres articulateurs, les nasales québécoises sont dipthonguées: le point de départ correspond à la voyelle orale puis on évolue en cours de nasale vers une plus grande antériorisation pour les antérieures et une plus grande postériorisation pour les postérieures (ainsi qu'une fermeture généralisée). Les nasales septentrionales ont une configuration orale différente (plus postérieure) des orales correspondantes tout au long de leur production, alors qu'en français méridional les timbres sont relativement équivalents pour orales et nasales. Ainsi les trois dialectes font varier les mêmes paramètres de façon différenciée, ce qui aboutit à un ensemble d'articulations covariantes à la fois cohérent du point de vue de leur structure interne et bien distinct des autres variétés de français. Ceci plaide en faveur de l'hypothèse d'une réponse adaptative de chaque dialecte aux contraintes phonétiques et phonologiques qui pèsent sur la nasalité vocalique.

BIBLIOGRAPHIE

[1] Wright, J.T. The Behavior of Nasalized Vowels in the Perceptual Vowel Space. *Experimental phonology.* Ohala & Jaeger (eds), New York, Academic Press, 45-67, 1986.

[2] Kawasaki, H. Phonetic explanation for phonological universals. *Experimental phonology,* 81-103, 1986.

[3] Ohala, J. J. & Busà, M. G. Nasal loss before voiceless fricatives: a perceptually-based sound change. *Rivista di Linguistica 7,* 125-144, 1995.

[4] Kingston, J. et Diehl, R. L. Phonetic Knowledge. *Language*, 70, 3, 419-453, 1994.

[5] Martin, P., Beaudoin-Bégin, A.M., Goulet M.J. & Roy J.P. Les voyelles nasales du Québec. *La linguistique*, 37, 2, Paris, PUF, 49-70, 2001.

[6] Gendron, J.-D. La méthode radiographique appliquée à la comparaison des articulations vocaliques en français canadien et parisien. *Proceedings 4th ICPhS*, 155-166, 1961.

[7] Delvaux, V. Etude aérodynamique de la nasalité en français. *Actes 23e Journées d'Etude sur la Parole*, 141-144, 2000.

TOUS NOS REMERCIEMENTS à Lucie Ménard ainsi qu'aux membres du Laboratoire de Phonétique de l'Uqàm.

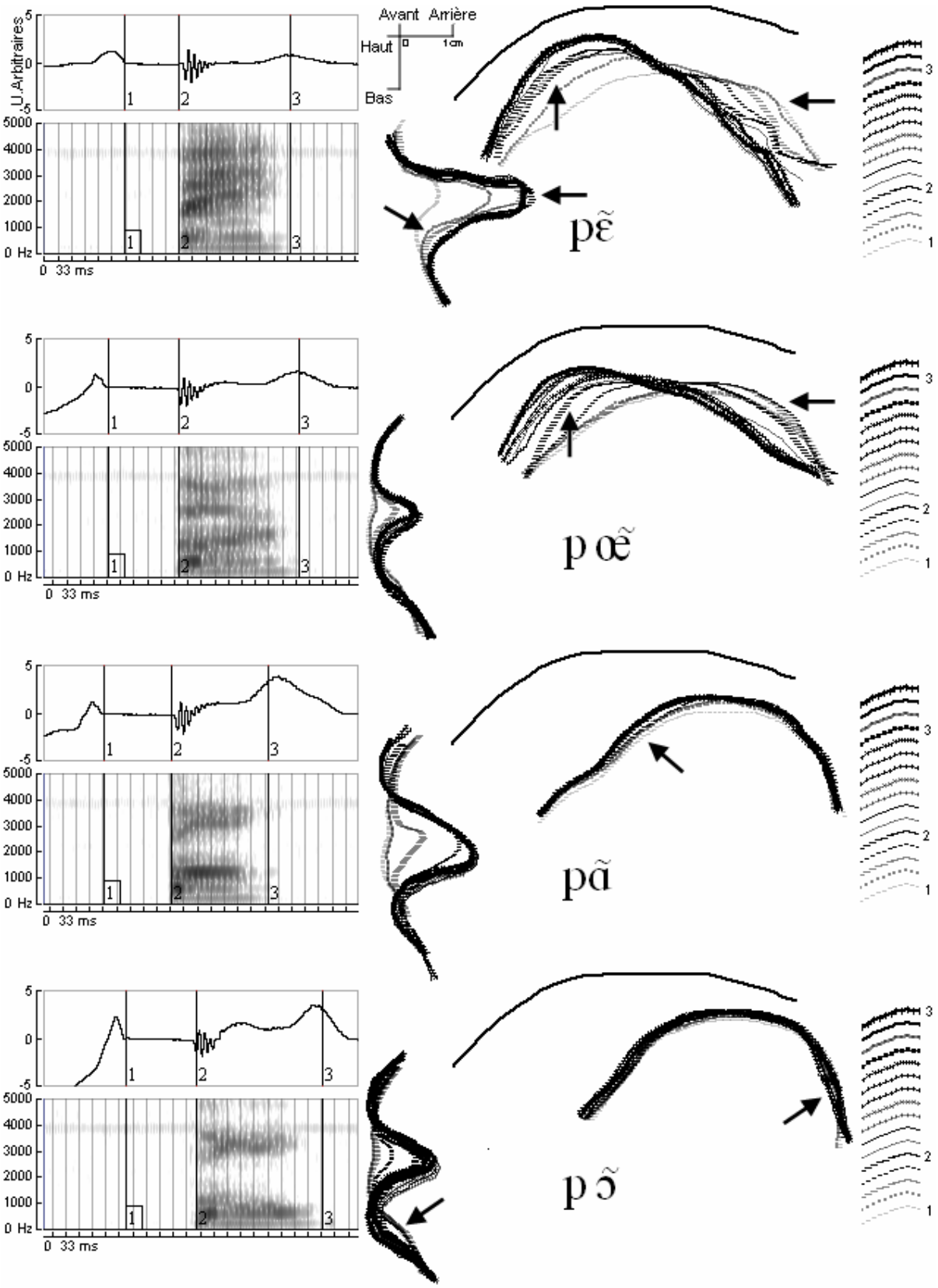


Figure 2 : De haut en bas, production de /pɛ̃, pœ̃, pɑ̃, pɔ̃/ par CE (Québec). A gauche: DAN, spectrogramme. A droite: contours des lèvres et contours de langue entre les repères temporels (1) et (3).

Session XIV

Conférence Invitée

Jeudi 15 juin 2006 - 09h00 10h00

De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux

Martine Adda-Decker

LIMSI-CNRS, Université de Paris-Sud, BP 133, 91403 Orsay CÉDEX, France
Mél : Martine.Adda@limsi.fr

ABSTRACT

This contribution aims at giving an overview of present automatic speech recognition in French highlighting typical transcription problems for this language. Explanations for errors can be partially obtained by examining the acoustics of the speech data. Such investigations however do not only inform about system and/or speech modeling limitations, but they also contribute to discover, describe and quantify specificities of spoken language as opposed to written language and speakers' speech performance. To automatically transcribe different speech genres (e.g. broadcast news *vs* conversations) specific acoustic corpora are used for training, suggesting that frequency of occurrence and acoustic realisations of phonemes vary significantly across genres. Some examples of corpus studies are presented describing phoneme frequencies and segment durations on different corpora. In the future large-scale corpus studies may contribute to increase our knowledge of spoken language as well as the performance of automatic processing.

1. INTRODUCTION

Contrairement à bien d'autres domaines de recherche autour de la parole, la reconnaissance automatique, qui s'effectue sur un flux acoustique continu, nécessite une modélisation de l'ensemble des phénomènes observés dans le signal : au-delà des mots auxquels est associée une représentation de type phonologique dans le dictionnaire de prononciation, il faut modéliser des respirations, des hésitations, des fragments de mots, des brouillons de parole peu ou pas articulés... Dans cette contribution nous allons faire d'abord un rapide état de l'art des systèmes de transcription automatique, présenter leurs performances et analyser les types d'erreurs les plus représentatifs. Nous allons ensuite poser la question de ce que peuvent nous apprendre ces erreurs de transcription. Ceci nous amène à utiliser progressivement les systèmes de transcription comme des instruments d'analyse de grands corpus oraux, d'abord sur les données utilisées pour l'apprentissage ou l'évaluation des systèmes, mais également sur des corpus à visée plus linguistique comme le corpus PFC (Phonologie du Français Contemporain) [15]. De telles études permettent par exemple de décrire et de quantifier des variantes de prononciations [3], des disfluences [9] et des réalisations acoustiques des sons [10]. Quelques adaptations méthodologiques des systèmes s'imposent afin de transformer un système de transcription en un instrument d'analyse de corpus oraux.

2. TRANSCRIPTION AUTOMATIQUE DE LA PAROLE

Il est admis que les progrès réalisés dans le domaine de la transcription ont été largement stimulés par les campagnes d'évaluation, lors desquelles les participants évaluent leur système sur un jeu de test commun et bénéficient de données d'apprentissage communes. Une première évaluation de systèmes de reconnaissance automatique multilingue, incluant les langues française, anglaise et allemande, a été mis en place dans le cadre d'un projet européen (LE-SQALE) en 1994 [26, 32]. Cette évaluation a permis de mettre en évidence pour la transcription automatique, des difficultés spécifiques aux différentes langues. Ainsi le français se caractérise par un nombre d'homophones particulièrement élevé en comparaison avec l'anglais et l'allemand [20]. La simple suite de phonèmes /la/ peut se transcrire suivant le contexte par les mots `la, là, l'a, l'as, las` et l'information acoustique ne permet pas de lever l'ambiguïté à elle seule. Une information sur le contexte de la séquence /la/ est nécessaire. Cette information de contexte est apportée dans les systèmes de transcription par les modèles de langage (des probabilités d'observation de mots et de séquences de mots). Ainsi des séquences comme `tu l'as vu`, `tu la retrouves` sont probables, en revanche des séquences comme `tu l'a` ou `tu las` sont fortement improbables.

Depuis une dizaine d'années, deux campagnes d'évaluation des systèmes de reconnaissance automatique en langue française ont eu lieu [13, 19] et permettent d'apprécier les progrès accomplis. La première campagne d'évaluation a été lancée sous l'égide de l'AUPELF-UREF il y un peu plus de dix ans. Il s'agissait ici de transcrire automatiquement des textes lus du journal *Le Monde* par des locuteurs différents et inconnus en utilisant le corpus BREF [25] enregistré au LIMSI autour des années 90. Ces recherches visaient à démontrer la capacité des systèmes à retrouver les mots prononcés à l'oral. Cette campagne a permis de montrer que des taux de mots erronnés proches de 10% pouvaient être obtenus. Il faut cependant garder à l'esprit que la parole lue est certes médialement de l'oral, mais qu'au fond elle reflète parfaitement la langue écrite. Ceci implique en particulier pour les systèmes, une relative facilité d'estimer des modèles de langage réalistes pour la tâche en question à partir de textes de journaux, dans la mesure où le signal acoustique correspond aux textes lus. Concernant les caractéristiques acoustico-phonétiques de BREF, la lecture à haute voix n'est pas une tâche quotidienne et entraîne pour la plupart des locuteurs une articulation plutôt lente et relativement soignée « qui colle à l'écrit ». Une telle prononciation respecte plutôt bien les hypothèses de modélisation acoustique des mots comme séquences de phonèmes qui est faite par les systèmes de transcription.

La deuxième campagne, ESTER (évaluation des systèmes de transcription enrichie d'émissions radiophoniques), financée par le programme interministériel français TECHNOLANGUE et organisée conjointement par l'AFCP, la DGA et ELDA vise un objectif beaucoup plus ambitieux, dans la mesure où il s'agit ici de parole journalistique

d'émissions radiophoniques, de différentes stations de radio. Même si une bonne partie des enregistrements correspondent à de la parole préparée, i.e. produite à partir d'une préparation écrite, elle est « convertie » à l'oral par des présentateurs et des présentatrices professionnels. Un pourcentage non négligeable des émissions correspond également à des interventions d'invités ou d'auditeurs, pour lesquels il y a souvent peu ou pas de préparation écrite. On est donc ici plus proche de la langue orale. Afin de permettre d'estimer des modèles de langage adaptés à l'oral journalistique (par opposition aux journaux écrits), la DGA, en collaboration avec le LIMSI, a entrepris la transcription manuelle de dizaines voire centaines d'heures de journaux radiodiffusés à la fin des années quatre-vingt dix. Dans cette dynamique a été développé le logiciel TRANSCRIBER [5], qui a trouvé un large succès pour la transcription de corpus oraux, bien au-delà de la communauté du traitement automatique de la parole. L'exercice de transcription manuelle pointe sur des problèmes qui se retrouveront également lors de la transcription automatique. Ainsi on peut se rendre compte à l'écoute attentive que bon nombre de mots sont souvent réalisés de manière incomplète. On pourrait être tenté d'écrire `ad'taleur` pour `à tout à l'heure`, tout comme on peut trouver fréquemment des « trucages orthographiques » comme `'ya` pour `il y a`. Nous avons préconisé pour les transcriptions manuelles le même principe de transcription en orthographe normative comme les conventions du GARS [6], avec un minimum d'indications de prononciations. Ce principe permet le mieux de converger vers des transcriptions stables indépendantes du transcripteur et avec un temps de transcription plus faible que si des annotations spécifiques étaient effectuées. Le problème de « trucages » des prononciations est cependant bien réel et nécessite des adaptations au niveau de la modélisation acoustique des mots.

Le passage de la lecture à la parole radiophonique a donc un impact au niveau des prononciations, avec des réalisations qui peuvent s'écarter de manière plus importante de prononciations canoniques [14]. Les mots outils fréquents, dont l'information est largement portée par le contexte sont souvent mal et peu articulés, et sous l'**effet de répétition** même des mots pleins thématiques (p.ex le mot `architecture` /aRʃitɛktyR/) bien prononcé pendant quatre, cinq fois en début d'émission, finit par être raccourci, ne préservant plus que certaines parties du mot, (comme par exemple [aRʃɛktyR]). Par opposition à une tâche de lecture sans auditoire, qui consiste alors à prononcer les mots écrits de manière relativement équitable, les émissions radiophoniques sont destinées à un large public dispersé et distant, et le souci de compréhension prévaut certainement ici à celui d'une simple articulation claire et équitable.

2.1. Quelques résultats de transcription

La reconnaissance de la parole, consiste à déterminer la meilleure suite de mots $\hat{m}$ à partir de l'observation acoustique x. Avec l'approche statistique ce problème repose alors sur la formule de Bayes

$$\hat{m} = \arg\max_m P(m/x) = \arg\max_m p(x/m)P(m)$$

Le décodeur doit mesurer la probabilité de toutes les suites de mots m possibles pour ce signal : $P(m/x)$. Le problème se transforme grâce à la formule de Bayes en une optimisation à deux termes $p(x/m)P(m)$ pour lesquels des modèles peuvent être estimés à partir de grands corpus d'apprentissage. Le premier terme $p(x/m)$ s'évalue grâce à des modèles acoustiques de mots, construits à partir de modèles acoustiques de phones via un dictionnaire de prononciation. Le deuxième terme $P(m)$ donne une estimation de la probabilité a priori de la séquence de mots m grâce aux modèles de langage N-grammes. La table 1 donne des ordres de grandeurs de quelques paramètres caractérisant les systèmes de transcription.

TAB. 1: Evolution des systèmes de transcription automatique du français : style de parole, données d'apprentissage et taux d'erreur de mots

		apprentissage		*%err.*
campagne	*style*	$p(x/m)$	$P(m)$	*mots*
AUPELF	lecture	BREF	*Le Monde*	12%
1996		100h	40M mots	
ESTER	journa-	radio	journaux, web	11%
2005	listique	100h	400M mots	

Pour les deux évaluations, effectuées sur des données de nature assez différente (lecture enregistrée en laboratoire, parole radiophonique grand public) les taux d'erreur sont proches de 10%. Dans certaines conditions et pour certains locuteurs professionnels les taux peuvent descendre autour de 5%. Mais il est actuellement difficile d'approcher des taux aussi faibles pour une population large de locuteurs. Pour les conversations téléphoniques en français, qui correspondent à un vrai genre oral, les taux d'erreur sont facilement supérieurs à 30% [4]. Certes les conditions acoustiques sont moins bonnes et contribuent à augmenter les erreurs, mais les problèmes essentiels pour la parole conversationnelle concernent à la fois l'estimation d'un modèle de langage approprié au genre traité, et les prononciations des mots avec la modélisation acoustique associée. Des problèmes supplémentaires concernent l'établissement d'une transcription de référence dans des zones disfluentes ou simplement mal articulées.

Par la suite nous allons nous limiter surtout à la parole radiophonique, pour laquelle on peut distinguer deux sources aux problèmes de transcription dans le cadre de la modélisation statistique exposé : est-ce que le système arrive à prédire de manière fiable les mots prononcés ? est-ce que les modèles acoustiques des mots reflètent les prononciations effectivement produites par les locuteurs ? Dans ce qui suit nous allons essayer d'analyser des erreurs de transcription en gardant en tête ces deux questions.

2.2. Analyse des erreurs

Le corpus ESTER dev04 [19] contient 10 heures de parole avec environ 94k mots dont un peu moins de 10k entrées lexicales distinctes. Sur ce jeu de données, le système du LIMSI [21] a produit un peu plus de onze mille erreurs dont sept mille substitutions, environ trois mille omissions et un peu moins de mille cinq cents insertions. Le taux de substitutions est un peu plus que le double des taux d'omissions, qui est lui-même le double du taux d'insertions. Dans l'analyse ci-dessous nous allons aborder plus en détail les erreurs sur les mots les plus fréquents, essayer de caractériser les mots bien ou mal reconnus et examiner la dispersion des erreurs dans le flot de parole via l'étude de zones d'erreur.

Est-ce que les mots fréquents sont bien reconnus ?

L'apparition des mots fréquents dans leur contexte est normalement bien apprise par le modèle de langage $P(m)$ à partir du corpus d'apprentissage. On peut donc s'attendre à un faible nombre d'erreurs dues simplement à un manque d'observation a priori. De même les mots fréquents à l'oral sont bien représentés dans les corpus audio et les modèles acoustiques doivent bien les représenter. Si les corpus ayant servi à l'estimation des modèles de langage et des modèles acoustiques représentent le même genre de données que le test, les mots fréquents devraient

être bien reconnus. Le taux d'erreur moyen sur les 20 mots les plus fréquents est de 10,5% qui est certes endessous du taux moyen de 12% d'erreurs obtenus sur les données de développement (11% pour l'évaluation), mais reste proche du taux d'erreur moyen. Comment expliquer les erreurs de reconnaissance des mots fréquents ?

La table 2 donne la liste des 20 mots (entrées lexicales) les plus fréquents dans le corpus de développement d'Ester. A gauche est donné le nombre d'occurrence de ces 20 mots, classés par rang de fréquence. A droite les mêmes 20 mots sont triés par taux d'erreur intra-classe décroissant (calculé pour chaque mot m_i comme le ratio des mots m_i mal reconnus, incluant substitutions, omissions et insertions de m_i par le nombre de mots m_i dans le corpus de référence). Le taux d'erreur entre parenthèses ne tient compte que des erreurs de substitution et d'omission. On peut voir clairement qu'il y a des tendances très différentes pour ces 20 mots les plus fréquents, qui expliquent à eux seuls plus d'un quart des erreurs commises (28%). Dans le tableau les taux d'erreur les plus élevés correspondent à des mots monophones, donc très courts où le modèle acoustique ne peut pas jouer un rôle discriminant important, et qui admettent en plus des homophones. L'homophonie implique que le choix du mot incombe au modèle de langage (dans l'hypothèse où le modèle acoustique a réussi). Ceci est le cas par exemple pour les paires (`et, est`) et (`à, a`) qui admettent des taux d'erreur autour de 20%. Mais à l'intérieur de chaque paire on peut observer une dissymétrie : `et` et `a` sont beaucoup plus facilement insérés, leur taux d'insertion correspond respectivement à 7,7 et 9,2%. Ceci s'explique par le modèle de langage qui pénalise plutôt l'insertion de `est` et de `à` face à leur contrepartie homophone. Le mot `il`, dont la prononciation canonique dans le système est /il/, a un taux d'erreur élevé de 18,8%. Or `il` est fréquemment réduit à [i] et admet ainsi un quasi-homophone `qui` au rang 19 et au moins deux homophones au-delà du rang 20 : `ils` et `y`, à des rangs supérieurs à 100. Le moins d'erreurs sont observées pour des formes plus longues (2 à 3 phonèmes) qui sont acoustiquement moins ambiguës. Les deux mots les plus fréquents `de` et `la` ont des taux d'erreur faibles de 6,7% et de 3,4%.

TAB. 2: Liste des 20 formes lexicales les plus fréquentes triées par leur nombre d'occurrences et triées par leur taux d'erreur intra-classe. Ce taux d'erreur tient compte des substitutions, omissions et insertions. Le chiffre entre parenthèses néglige les insertions.

forme	*#occ*	*rang*	*forme*	*%err (-%ins)*	*rang*
de	5355	1	et	25,4 (17,7)	4
la	2684	2	est	20,0 (17,1)	8
le	3011	3	a	19,5 (10,3)	14
et	1927	4	il	18,8 (16,2)	15
à	1887	5	à	15,6 (10,2)	5
l'	1840	6	un	13,1 (9,6)	11
les	1800	7	que	9,8 (7,6)	16
est	1367	8	qui	9,6 (7,0)	19
des	1378	9	en	9,6 (7,3)	10
en	1315	10	l'	9,5 (8,3)	6
un	1311	11	les	9,0 (8,3)	7
d'	1116	12	le	8,7 (6,2)	3
du	1101	13	des	8,5 (7,4)	9
a	1815	14	d'	7,8 (6,5)	12
il	916	15	de	6,7 (3,9)	1
que	913	16	une	5,8 (4,3)	18
pour	882	17	dans	5,0 (4,6)	20
une	790	18	pour	4,4 (2,2)	17
qui	797	19	du	4,3 (3,6)	13
dans	724	20	la	3,4 (2,4)	2

En résumé pour les mots fréquents les taux d'erreur au-delà du taux d'erreur moyen s'expliquent essentiellement par deux facteurs : formes courtes et homophonie. Pour les couples de mot (`et, est`) et (`à, a`) particulièrement problématiques, il peut être intéressant au niveau des paramètres acoustiques du système, d'introduire une information prosodique (en particulier l'évolution de la fréquence fondamentale) contenant éventuellement quelques marques distinctives, pour le moment négligées, ainsi qu'un post-traitement morpho-syntaxique spécifique afin de faire progresser les performances. De manière plus générale informations prosodiques et morpho-syntaxiques devraient contribuer à la précision des systèmes de transcription dans le futur.

Quels mots sont les mieux/les moins bien reconnus ?

En examinant les erreurs par mot on se rend compte que la formule

$$\%err = \frac{(sub(m_i^r) + del(m_i^r) + ins(m_i)) * 100}{occ(m_i^r)} \quad (1)$$

n'inclut pas dans sa mesure si le mot m_i est facilement substitué à la place d'un autre mot m_j. Si on veut examiner les causes d'erreur, il peut être important d'inclure dans la mesure les deux types de substitution : le premier type est celui de la formule classique ci-dessus, où le mot m_i de la référence est substitué par le mot m_j de l'hypothèse. Le deuxième type de substitution correspond à la situation inverse où m_j de la référence est faussement transcrit comme m_i dans l'hypothèse. On peut définir une nouvelle mesure de substitutions :

$$sub'(m_i^r) = \frac{\sum_{j \neq i} sub(m_j^h, m_i^r) + \sum_{j \neq i} sub(m_i^h, m_j^r)}{2} \quad (2)$$

ce qui donne la formule suivante :

$$\%err' = \frac{(sub'(m_i^r) + del(m_i^r) + ins(m_i))}{occ(m_i^r)} \quad (3)$$

Le tableau 3 montre des exemples de mots pour lesquels le taux d'erreur augmente beaucoup en passant de la formule classique (*%err*) à la formule modifiée (*%err'*). Ces mots se retrouvent ainsi souvent faussement dans l'hypothèse. La première ligne montre qu'il n'y a pas d'erreur de type *%err* pour le mot `membres` sur les 39 occurrences dans le corpus de référence. Toutes les erreurs concernant le mot `membres` se produisent dans l'hypothèse et sont des erreurs de substitution pour le mot `membre` dans la référence : les probabilités du modèle de langage favorisent la forme au pluriel. Dans un grand nombre de cas le taux *%err'* plus élevé s'explique par des homophones morphosyntaxiques dans la référence, dont les probabilités d'émissions sont plus faibles dans le modèle de langage $P(m)$. Pour le mot `eh` l'explication est une incohérence au niveau des conventions de transcription (`et bien` *vs* `eh bien`) et pour les mots `oui` et `non` pour lesquels les deux taux sont très élevés, les erreurs sont majoritairement des insertions et omissions. Il reste encore des progrès à accomplir pour assurer des taux d'erreur faibles pour des mots importants comme `oui` et `non`. Ces derniers, typiques d'interactions orales, sont relativement peu représentés dans les corpus d'apprentissage. Il est intéressant de voir que `Paris` peut apparaître facilement à la place du mot de référence `pays`. Cette confusion est certainement imputable au modèle de langage, une explication au niveau des modèles acoustiques suggérerait que les mots `Paris, parisien...` peuvent se prononcer de manière proche de `pays`.

Il y a beaucoup plus de formes lexicales pour lesquelles la situation est l'inverse avec un taux %err plus élevé que le taux %err'. En effet tous les mots plutôt rares dans le corpus d'apprentissage, qui admettent un (quasi)-homophone

TAB. 3: Exemples de mots avec *%err < %err'*. Ces mots se retrouvent de manière erronnée dans l'hypothèse. La dernière colonne indique si les erreurs concernent plutôt omissions (del), insertions (ins) ou substitutions (exemples de mots).

forme	*%err*	*%err'*	*#occ.*	*comment.*
Paris	2.8	7.7	71	pays
membres	0	9.0	39	membre
jour	2.6	10.5	38	jours, jouera
reste	2.8	11.1	36	restent
rencontre	2.9	8.6	35	rencontr-es,ent,er
pourrait	3.8	13.5	26	pourraient
cette	3.9	10.2	230	sept, cet, ces
était	16.4	20.8	158	étaient, été, est
non	25.5	35.1	47	del ont
eh	35.3	42.6	34	et bien → eh bien
oui	42.4	48.3	59	del/ins

fréquent ne seront pas facilement proposés à tort par le système de transcription automatique. La table 4 donne quelques exemples.

TAB. 4: Exemples de mots avec *%err > %err'*. Ces mots sont facilement substitués lors du décodage et n'apparaissent que rarement de manière erronnée dans l'hypothèse. La dernière colonne indique si les erreurs concernent plutôt omissions (del), insertions (ins) ou substitutions (exemples de mots).

forme	*%err*	*%err'*	*#occ.*	*comment.*
George	13.0	6.5	23	Georges
Abbas	41.0	21.8	39	Abbass, baisse
Al	51.0	33.0	47	del a, à, Alma
responsables	13.0	6.5	23	responsable
officielle	15.4	7.7	39	officiel
mettre	16.1	8.1	31	mais, promet
cour	18.6	11.6	43	cours, recours
ceux	22.5	15.4	40	del ce
êtes	24.1	15.5	29	est
eu	30.2	21.4	63	eus, vu
ai	40.7	32.7	55	del est
ils	42.8	31.1	201	il, qui
elles	56.7	38.3	30	elle
bon	37.2	26.7	43	mon, ben
me	53.6	46.4	28	del
hein	96.4	76.8	28	del

On peut remarquer que les taux d'erreur sont particulièrement élevés dans la table 4. En utilisant la mesure classique %err on trouve environ 2500 entrées lexicales avec un taux d'erreur de mot supérieur à 25% (8% du corpus et 40% des erreurs). Il s'agit ici, comme nous l'avons déjà vu ci-dessus, soit de mots outils admettant des (quasi)-homophones à fréquence plus élevée. Dans les mots à taux d'erreur très élevés on trouve des mots spécifiques aux discussions orales, comme oui, écoutez, savez, quoi, ben, ah, là, j', ai, suis, , , des mots pleins courts et donc acoustiquement difficile à discriminer comme gens, air, eau, or, sports.

Pour finir sur une note plus positive on peut trouver presque 6000 entrées lexicales parfaitement reconnues (taux d'erreur=0%). Parmi ces mots qui représentent 20% du corpus, on peut trouver aussi bien des mots outils, des mots pleins et des noms propres, comme aujourd'hui, toujours, lors, selon, plusieurs, notamment, soixante, gouvernement, syndicats, secrétaire, membres, coopération, national, Jean, Washington, Bagdad, Pakistan, Rabat, ONU...

On peut remarquer qu'un mot court et acoustiquement difficile comme lors est parfaitement reconnu. Il s'agit ici d'un mot fréquemment utilisé dans le style journalistique et donc favorisé par le modèle de langage. Ceci se fait alors au détriment d'homophones moins fréquent comme le montre l'exemple suivant [1] :

```
REF : sur la tombe d' andré breton il est écrit
je cherche L' OR du temps
HYP : sur la tombe d' andré breton il est écrit
je cherche ** LORS du temps
```

Le pourcentage d'erreurs dû aux homophones morphosyntaxiques (comme membres, membre) est moins important lors de l'évaluation ESTER en 2004 (inférieur à 20%) que pour celle d'AUPELF en 1996 (autour de 30%). Alors que des modèles de langage incluant une information morphosyntaxique n'ont permis d'améliorer le taux d'erreur que de 0,1% (absolu), le fait d'utiliser des modèles de langage incluant un plus grand nombre de trigrammes diminue naturellement les erreurs. La décision du choix entre différents homophones peut se faire plus souvent en tenant compte du contexte plutôt que de faire appel au mécanisme de repli éliminant l'information contextuelle.

Est-ce-que les erreurs se produisent de manière isolée ou en groupe ?

Alors que dans la partie précedente, l'analyse portait sur les mots pris de manière isolée, nous allons examiner ici les erreurs telles qu'elles se produisent dans le flot de parole continue. On peut se poser la question si les erreurs arrivent plutôt de façon isolée ou si une erreur en entraîne d'autres dans son voisinage immédiat, dans la mesure où une erreur risque d'engager le modèle de langage sur une fausse piste pour ses prédictions. Pour éclairer cette question nous avons compté, en plus du nombre d'erreurs, le nombre de zones erronnées dans les transcriptions automatiques du corpus de développement de ESTER 2004, une zone étant définie comme une suite d'erreurs consécutives, les erreurs pouvant être de trois types : substitution, insertion et omission. La table 5 donne un exemple de zone d'erreur de longueur 5. Pour ces dernières il y a 131 de telles zones, qui contribuent avec 655 erreurs à 6% du taux d'erreur.

TAB. 5: Exemple de zone d'erreur de longueur 5. S : substitution, O : omission

REF :	l'	AGGRAVE	ET	PEUT	LE	TUER
HYP :	l'	AGGRAVER		PAUL		TUÉS
ERR :	-	S	O	S	O	S
comm. :		homophone		/pølə/ → [pɔl]		hom.

La figure 1 donne la distribution des zones d'erreur et le nombre d'erreurs par zone en fonction de la longueur des zones. Cette distribution suit en gros une loi de Zipf avec un très grand nombre de zones erronnées de longueur 1 et très peu de zones de longueur élevée. La courbe montre que dans 4000 cas la zone est de longueur 1 et n'entraîne donc pas de dommages collatéraux. Nous avons examiné les types d'erreurs (substitutions, omissions, insertions) en fonction de la longueur de la zone. Il y a environ 65% de substitutions, un peu plus de 20% d'omissions et un peu plus de 10% d'insertions avec des taux de substitutions plus élevés pour les zones de faible longueur et plus faibles pour les zones de longueur élevée. Dans ce dernier cas les taux d'omissions peuvent devenir particulièrement importants.

[1] Alors que le système produit une transcription respectant la casse, la mesure des erreurs est insensible à la casse : tous les mots bien reconnus sont en minuscules, les substitutions sont en majuscules et les insertions/omissions sont marquées par des étoiles. REF, HYP indiquent la transcription manuelle de référence ainsi que la meilleure hypothèse de

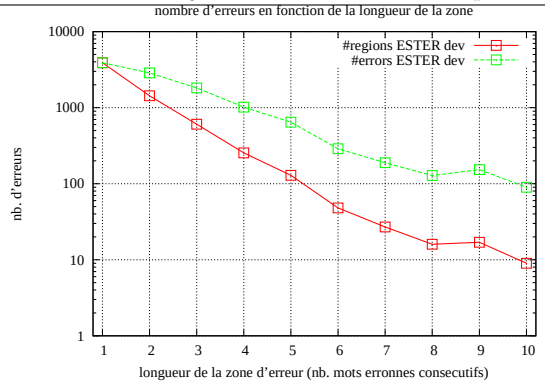


FIG. 1: Nombre d'erreurs et de zones d'erreur en fonction de la longueur de la zone sur le corpus de développement d'ESTER 2004. La longueur d'une zone d'erreur est définie comme le nombre maximum d'erreurs consecutives (incluant substitution, insertion et omission).

L'exemple suivant, avec 3 zones d'erreurs (2 zones de longueur 1 et 1 de longueur 4) permet d'illustrer quelques cas-types d'erreurs.

```
REF : ELLE atteindront sur le bassin parisien
PRÈS DE LA LOIRE    sur le sud de la bretagne la
vendée vingt trois À vingt quatre degrés
HYP : ELLES atteindront sur le bassin parisien
** ** ** PRÉVALOIR  sur le sud de la bretagne la
vendée vingt trois * vingt quatre degrés
```

La première zone d'erreur de longueur 1 correspond en fait à une erreur de la transcription de référence et le système rétablit ici une erreur dans la référence [2]. On peut estimer autour de 2% le taux d'erreur résiduel humain. Ce taux varie évidemment en fonction du soin et du coût dépensés pour la production de la transcription de référence. La zone d'erreur de longueur quatre implique un syntagme prépositionnel `près de la Loire`, dont le spectrogramme est montré dans la figure 2. L'hypothèse `prévaloir` à trois syllabes peut être considérée comme spectralement et surtout temporellement acceptable, étant donnée la réalisation du syntagme en question sous forme de trois syllabes sous une forme comme [pRɛ.dla.lwaR] ou [pRɛ.la.lwaR]. La suite de mots outils `près de la` est articulée rapidement en 300 ms (environ 50 ms pour le mot `de`), la même durée que le noyau du syntagme `Loire`. La dernière zone d'erreur correspond à une élision du mot outil (monophonème) `à` dans un contexte vocalique gauche identique (`trois /tRwa/`). La durée mesurée sur la séquence /aa/ enchaînée est inférieure à 100 ms. Ceci est une situation particulièrement favorable aux élisions si les deux voyelles sont enchaînées sans césure [3].

Nous terminerons cette partie par une sélection d'extraits montrant différentes situations d'erreurs. Ces exemples illustrent que les décisions erronnées du système n'ont en général pas d'explication simple, dans la mesure où la décision est prise en fonction de la réalisation du locuteur en appliquant conjointement modèles acoustiques et modèle de langage.

Dans l'exemple suivant on trouve le même type d'erreur de suppression de voyelle que précédemment, dans la séquence `déjà à`. La liaison entre `elles` et `appellent`, bien que autorisée par le système grâce à

transcription automatique respectivement.

[2] La forme `elles` arrive à être reconnu face à son homophone plus fréquent `elle`, car le mot suivant `atteindront` marque acoustiquement le pluriel et le trigramme `elles atteindront sur` existe dans le modèle de langage.

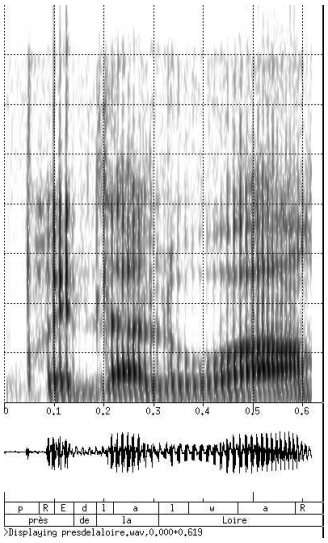


FIG. 2: Spectrogramme de la suite `près de la Loire` reconnue comme `prévaloir`. On observe une chute presque complète du mot `de`, esquissé uniquement par l'occlusion voisée du /d/, la phase d'explosion correspond au /l/ de `la`.

un phonème /z/ optionnel, provoque une zone d'erreur de longueur 3 qu'on doit attribuer plutôt au modèle de langage (cf. erreur `elles, elle` dans la table 4) qu'à un problème de modélisation acoustique.

```
REF : ELLES ** APPELLENT déjà À DES MOBILISATIONS
dans le public
HYP : ELLE S' APPELLE déjà * *** DÉMOBILISATION
dans le public
```

L'exemple suivant illustre la différence de précision de transcription automatique que l'on peut observer sur des mots qui ponctuent l'oral (ligateurs, incises...) : `alors, hein, je le signale`. On peut remarquer que l'hypothèse produite est significativement plus courte, ce qui reflète à l'écrit la réalisation acoustique contractée, réduite de ces mots ou suites de mot.

```
REF : ALORS françois chérèque HEIN JE le SIGNALE
qui sera l' invité de
HYP : À françois chérèque **** ** le SIGNAL qui
sera l' invité de
```

Comme déjà évoqué dans l'introduction, l'effet de répétition entraîne de la part des locuteurs, des performances de prononciation éloignée des prononciations standard, canoniques. Ci-après une réalisation contractée de `ministre de`, reconnue comme `minimiser`. Cette erreur est clairement d'origine acoustique, le modèle de langage favorisant sans ambiguïté le `ministre de l'intérieur`.

```
REF  :  nicolas sarkozy MINISTRE DE l' intérieur
bonjour
HYP : nicolas sarkozy *** MINIMISER l' intérieur
bonjour
```

Ici encore une erreur d'origine plutôt acoustique : la contraction de `c'est les` en `ses` dans l'hypothèse met en évidence une articulation rapide et peu précise sur ces suites de mots outils.

```
REF : pourtant et C' EST LE paradoxe les nouveaux
moyens
HYP : pourtant et * ** SES paradoxe les nouveaux
moyens
```

L'analyse des erreurs de transcription montre que bon nombre d'erreurs sont simplement dus aux prédictions

insuffisantes des modèles de langage. En particulier dès lors que plusieurs homophones sont en compétition, le choix ne repose plus que sur le modèle de langage. En revanche sur les erreurs où des prononciations différentes sont proposées entre référence et hypothèse, la modélisation acoustique influe sur la décision. Examinant différents types de parole (présentations journalistiques ici, à comparer d'un côté avec la lecture et d'un autre côté avec des conversations) et en écoutant le signal sur les parties erronnées comme sur des parties bien transcrites, il apparaît que les prononciations varient beaucoup en fonction du style de parole. Ces variations n'impliquent pas nécessairement des erreurs. En effet les modèles acoustiques de phones en contexte intègrent de manière implicite une grande partie des variantes [1]. En passant de la lecture à des conversations privées il est cependant nécessaire d'ajouter davantage de variantes dans le dictionnaire de prononciation. Il s'agit essentiellement de l'adjonction de formes contractées par rapport à la forme canonique (pleine) pour les mots outils très fréquents, pour les nombres, les dates et autres locutions. De nombreuses erreurs de transcription impliquent encore le schwa final comme dans `de, le, je, ne, se, ce, que, te, me`, dont la chute n'est pas prévu pour l'instant devant consonne. Or dans ces mots CV monosyllabiques à noyau faible le schwa tombe très facilement dès lors que le mot précédant se termine par une voyelle. Ainsi dans la séquence `près de la` le schwa du `de` est fréquemment omis à cause de la voyelle finale /ɛ/ du mot `près` qui précède. D'autres voyelles, consonnes, clusters voire des syllabes entières peuvent disparaître ou les syllabes peuvent se restructurer. Il est connu que les liquides sont facilement élidées ainsi que la consonne /v/. Ces observations amènent naturellement à se poser des questions sur les régularités d'apparition de ces variantes de prononciations et de manière plus générale sur les spécificités de l'oral par rapport à l'écrit oralisé, l'usage des mots, des tournures de phrases, des hésitations et autres disfluences. Il est important d'essayer de tirer profit des outils de transcription automatique et des grands corpus accumulés pour l'apprentissage des modèles du système afin d'en dégager des nouvelles connaissances sur l'oral.

3. ANALYSE DE CORPUS

Nous abordons ici l'analyse de corpus oraux en gardant parmi les objectifs de mieux cerner ce qui peut poser problème aux systèmes de transcription automatique. Nous avons souligné le problème de réalisations acoustiques réduites et contractées, provoquant des erreurs de transcription, et nous avons évoqué le besoin d'introduire des variantes de prononciations en fonction du style de parole. Ces problématiques sont convergentes à des objectifs de recherche en phonétique et phonologie, comme l'étude de la variation phonologique en français, en particulier la liaison et le schwa [12, 17, 16]. Des premiers travaux quantitatifs et qualitatifs sur la réalisation du schwa et de la liaison [8] ont déjà été effectués sur grands corpus, travaux qui viennent compléter des études plus fines sur corpus contrôlés [18, 28]. Plus récemment nous nous sommes intéressés aux restructurations syllabiques en parole spontanée faiblement contrôlée, aux disfluences dans le même type de corpus, aux hésitations en différentes langues [30, 10]. Des études des triangles vocaliques à partir de grands corpus ont montré l'impact de la durée sur la réalisation des voyelles, montrant un mouvement centripète des valeurs formantiques pour des durées de segment décroissantes. Ce résultat est à mettre en relation avec la discussion des erreurs de transcription présentée dans la section précédente et l'étude de la durée des segments ci-dessous. Cette tendance semble être indépendante de la langue [22]. Des études des variantes de prononciations en fonction de l'accent régional sont en cours dans la communauté autour du projet PFC (Phonologie du Français Contemporain) [15] avec un corpus visant à réunir quelques centaines d'heures de variétés de français en différents styles.

Dans ce qui suit nous allons utiliser différents corpus pour aborder la question de la fréquence des phonèmes pour la langue française [11, 27, 7] sur de grands corpus, et vérifier s'il y a des variations en fonction du genre de corpus. Des distributions contextuelles seront ensuite présentées. En effet pour la transcription automatique les fréquences des phonèmes, et en particulier les fréquences de phonèmes en contexte sont exploitées pour la modélisation acoustique. Ainsi le passage de modèles acoustiques de phones hors contexte à des modèles triphones (tenant compte des phonèmes gauche et droit) entraîne un gain significatif en précision de transcription. Nous présentons ensuite une étude comparative des durées segmentales entre genres de corpus et, à la fin une étude prosodique qui malgré son statut préliminaire permet d'envisager un large éventail de travaux à base de corpus étiquetés.

3.1. Corpus utilisés

Dans les études présentées ci-après nous faisons appel à trois corpus : un corpus de parole journalistique de 25 heures provenant de différentes stations de radios (*France Inter, France Infos*) et de chaînes de télévision (*France2, France3*), un corpus de parole conversationnelle par téléphone (entre amis, membres d'une même famille, ainsi qu'entre inconnus) et une partie du corpus PFC (phonologie du français contemporain), incluant différents styles de parole, avec de la lecture (liste de mots et texte) et des entretiens entre connaissances. Pour PFC nous considérons 12 points d'enquête, notamment Aveyron-Paris [3], Biarritz, Brécey, Brunoy, Dijon, Douzens, Lacaune, Lyon-Villeurbanne, Nyon, Roanne, Rodez et Vendée. La parole journalistique est caractérisée par le fait qu'il s'agit de parole préparée et publique, s'adressant à un public hétérogène n'interagissant pas (ou très peu) avec le locuteur. Le locuteur a souvent le monopole de la parole. Pour les conversations téléphoniques, il s'agit de dialogues privés où les interlocuteurs négocient leur tour de parole, mais ne disposent que du canal audio pour la communication. La situation téléphonique très naturelle fait que les interlocuteurs oublient facilement que la parole est enregistrée. Le corpus PFC contient de la parole lue et des conversations entre deux ou plusieurs personnes d'un même cercle de connaissances autour d'un micro. La durée et des facteurs caractérisant la production de l'oral sont donnés dans la table 6 pour les différents corpus.

TAB. 6: Caractéristiques des corpus oraux examinés.

corpus	*durée*	*parole/ interaction*	*auditoire*
radio TV	25h	préparée $\sim$ monologue	large public
convers. tél.	120h	spontané dialogue	1 ami
PFC	32h	lu+spont.	$<= 3$ pers
- mots	6h	lecture monologue	-
- texte	5h	lecture monologue	-
- entretien	21h	spontané dialogue	$<= 3$ pers

[3] Le point d'enquête Aveyron-Paris regroupe des locuteurs aveyronnais vivant depuis de nombreuses années à Paris.

3.2. Fréquence des phonèmes

Les corpus ont été alignés phonémiquement par le système de transcription automatique. Le dictionnaire de prononciation ne contient que des prononciations canonique incluant liaisons et schwa optionnels. Les fréquences des phonèmes sont ainsi calculées sur les trois corpus. La figure 3 donne les pourcentages des voyelles dans les corpus. On peut voir que les trois types de corpus suivent globalement la même évolution. Nous avons pris la courbe du corpus journalistique (en rouge) comme référence, pour laquelle nous avons indiqué les pourcentages des voyelles sur l'axe Y. Sur les axes X et Y, le schwa /ə/ et la voyelle centrale ouverte /œ/ sont codés ensemble par le symbole `x` ; la voyelle centrale fermée /ø/ est codée `eu`. Les deux /o/ ouvert et fermé sont comptés ensemble (o,c) et occupent ainsi 3,6% du corpus, chacun représentant environ 1,8%. Pour PFC nous donnons une courbe globale intégrant les divers types de parole de 12 points d'enquête. On peut observer que pour les conversations téléphoniques on a coté voyelles antérieures fermées, 1% (absolu) de /e/ en moins par rapport à la parole journalistique et pour les voyelles ouvertes, environ 2% de /ɛ/ et 2% de /a/ en plus, ainsi qu'un peu moins de 1% de /o,ɔ/ en moins. Pour les voyelles nasales il y a surtout un petit déficit pour la voyelle /ã/ dans les conversations téléphoniques. La courbe PFC reste proche des deux autres courbes et nous allons examiner les fréquences d'occurrences ici en fonction des différents styles (voir figure 5). Les entretiens PFC (guidés et libres) sont comparés aux conversations téléphoniques dans la figure 5 (à gauche). Deux courbes supplémentaires sont rajoutées afin de comparer quelques points du Nord (Brécey, Brunoy, Vendée, Dijon) à deux points du Sud (Douzens, Lacaune). Contrairement aux conversations téléphoniques, les entretiens PFC ont des taux de /e/ et de /ɛ/ sensiblement identiques. Il y a environ 2% de schwa supplémentaires dans les conversations téléphoniques que dans les entretiens PFC ; la courbe PFC-Sud (Douzens et Lacaune) a environ 2% de schwa en plus que la courbe PFC-Nord (4 points d'enquête). Dans la figure 5 à droite nous comparons la courbe PFC globale au sous-ensemble formé par le texte lu. Il est intéressant de noter que la courbe du texte s'écarte significativement des autres courbes par un déficit de /a/ (plus de 2% en absolu). En revanche il y a un nombre important de schwa, qui fait passer légèrement le schwa devant le /a/, le schwa devenant ainsi le phonème le plus fréquent. Ceci est dû à la forte présence de schwas dans le sud, qui est une des marques de l'accent méridional. Il faut cependant garder à l'esprit que le texte PFC *Le maire de Beaulieu* a été construit entre autre pour l'étude du schwa dans le français régional. Cette construction a eu un effet non négligeable sur la distribution des voyelles.

La figure 4 montre le pourcentage des consonnes du français observés dans les trois corpus. Globalement les courbes d'occurrence des consonnes suivent la même évolution sur les différents genres de corpus. Les consonnes les plus fréquentes du français sont /R/, /l/, /s/, /t/. On voit que le classement varie légèrement en fonction des genres de corpus examinés. Concernant la parole spontanée on peut voir que la parole téléphonique génère des proportions de /m/ et de /w/ plus élevées, ce qui est largement dû, pour le /m/ à une proportion élevée de mots comme `mais, moi, me` et des interventions de « backchannel » `hum` particulièrement élevé ici. Le /w/ provient de nombreuses occurrences de `oui, ouais, moi, voilà, toi, vois, crois`. On peut remarquer que ces mêmes mots enrichissent les statistiques du /a/ et du /ɛ/, comme nous avons pu le constater côté voyelles.

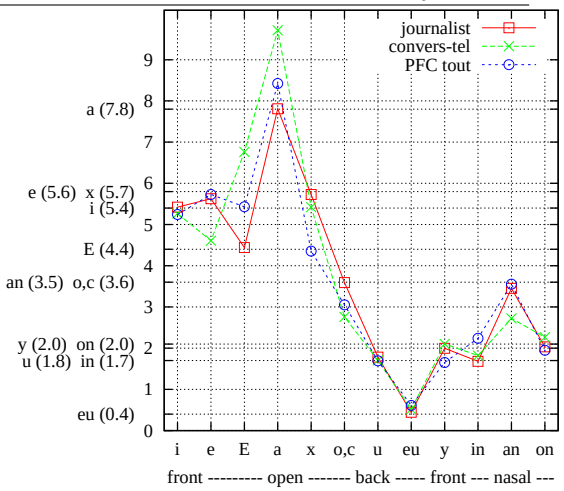


FIG. 3: Pourcentage d'occurrence des voyelles sur 3 types de corpus **journalistique, conversations téléphoniques, PFC divers, PFC**. Sur l'axe Y sont reportées les voyelles avec leur pourcentage du corpus journalistique, établissant ainsi une échelle de classement pour les voyelles.

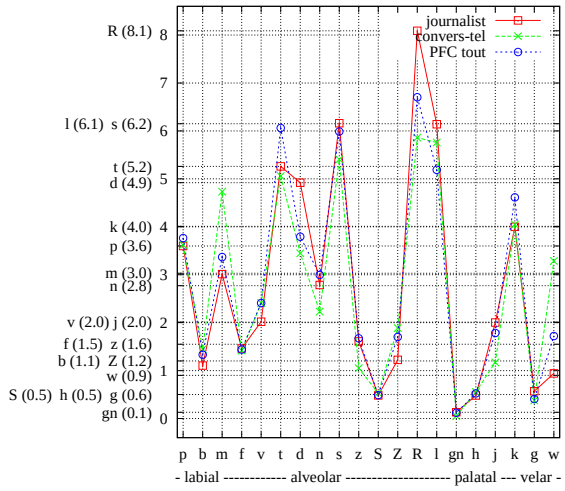


FIG. 4: Pourcentage d'occurrence des consonnes données sous forme de pourcentage de corpus. Sur l'axe Y l'échelle de classement des consonnes suit le corpus journalistique.

3.3. Fréquence d'occurrence des voyelles en contexte

Les fréquence d'occurrence des voyelles en contexte ont été mesurées sur les trois types de corpus : journalistique, conversations téléphoniques et PFC. Par manque de place nous ne présenterons ici que les mesures pour le corpus journalistique (voir Figure 10). Concernant les contextes nous avons distingué 6 classes consonantiques mélangeant pour ces classes des critères articulatoires, de fréquence d'occurrence et d'économie de présentation. Nous avons utilisé les lieux d'articulation tels qu'ils sont décrits dans le tableau synthétique des consonnes de l'IPA et nous avons considéré les liquides /l/ et /R/ très fréquentes en français séparément . A chaque classe est associé un code couleur dans les histogrammes : labial (/p/, /b/, /m/, /f/, /v/) en rouge, alvéolaire (/t/, /d/, /n/, /s/, /z/) en vert, postalvéolaire (/ʃ/, /ʒ/) en bleu foncé, liquide /R/ en rose, liquide /l/ en bleu ciel et palato-vélaire (/gn/, /ɥ/, /j/, /k/, /g/, /w/) en jaune. La dernière classe de consonnes palato-vélaires a été motivée surtout par une volonté de limiter le nombre de classes à six, pour des raisons de présentation des résul-

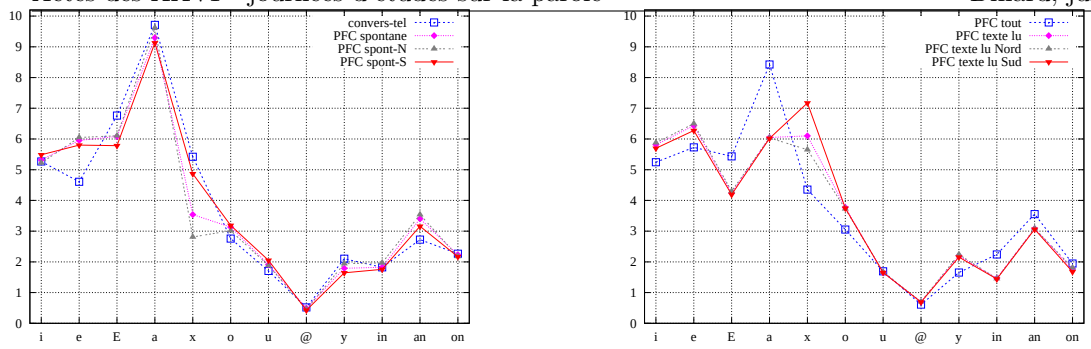


FIG. 5: Fréquences d'occurrence des voyelles. **A gauche :** parole spontanée :conversations téléphoniques et entretiens PFC. **A droite :** corpus PFC (tout et parole lue) : les fréquences d'occurrence des voyelles sont obtenues sur 12 points d'enquête. Pour le texte lu 2 courbes sont rajoutées comparant quelques points du Nord (Brécey, Brunoy, Vendée, Dijon) à deux points du Sud (Douzens, Lacaune).

tats. Ainsi pour un corpus donné les mesures tiennent sur une page sous forme d'un jeu de 6 histogrammes pour les 6 contextes consonantiques gauches. Ensuite dans chaque figure, ayant un contexte gauche fixé, on peut voir pour chaque voyelle la fréquence d'occurrence de chacune des 6 classes de contextes consonantiques droits. Ici les fréquences d'occurrences correspondent aux comptes dans les corpus sans ramener les comptes aux pourcentages d'occurrence.

Sans vouloir trop commenter ces mesures brutes, on peut faire quelques observations. Le contexte gauche alvéolaire (histogramme du milieu à gauche) est le plus riche en occurrences et le plus varié en contextes droits. En particulier on peut observer que cette classe apparaît fréquemment en même temps à gauche et à droite (contributions vertes importantes dans les barres des histogrammes), alors que ceci est beaucoup moins vrai pour les autres classes (par exemple peu de rouge dans la première figure en haut à gauche pour le contexte gauche labial, peu de rose dans la figure en haut à droite pour les contextes gauches de /R/ etc.). Contrairement aux données journalistiques, les conversations téléphoniques ont les fréquences les plus élevées pour le contexte gauche labial et les voyelles ouvertes /a/ et /ɛ/, ce qui rejoint nos observations sur les phonèmes hors contexte. Pour le corpus journalistique, le contexte gauche alvéolaire est le plus productif pour pratiquement toutes les voyelles, mise à part le /u/ plus fourni pour le contexte gauche labial, et la voyelle nasale /ɔ̃/ qui émerge particulièrement dans le contexte gauche palato-vélaire avec les deux consonnes /j/ et /k/ (avec des mots comme `question, millions, région, contre, compte, conseil`). Pour les 3 corpus (journal, conversations, PFC) on note des comptes élevés de la voyelle /a/ en contexte palato-vélaire gauche. Ceci est dû aux nombres comme `trois, quatre, quarante, soixante` et aux mots fréquents comme `quoi, crois, voilà, soir, avoir`. Sur les 18k contextes *palato-vélaire*-/a/ répertoriés pour le corpus journalistique, 8k proviennent de contextes /w/-/a/. Comme pour les fréquences des phonèmes, nous trouvons ici des différences assez marquées entre genres de corpus, variations qui sont à mettre en lien avec les variations du lexique. Des études plus fines sont nécessaires dans le futur pour compléter ces premières mesures présentées ici. Elles illustrent cependant que le rang de fréquence des voyelles varie en fonction du contexte phonémique gauche et droit.

3.4. Durées

Si nous pouvons mesurer des différences importantes de durées sur les segments phonémiques entre corpus de différents styles, ceci suggère qu'il existe bel et bien des différences significatives entre les prononciations qui contribuent à dégrader les performances lors de la transcription automatique. A partir de l'alignement automatique avec des dictionnaires de prononciation standard (incluant peu de variantes) la figure 6 donne la distribution des segments phonémiques en fonction de leur durée pour les corpus journalistique et de conversations téléphoniques. Plus le maximum de la distribution se trouve décalé vers la gauche (i.e. vers les segments courts) plus la courbe indique un risque de désaccord entre prononciations standard attendus et prononciations effectivement réalisées par les locuteurs, ces réalisations présentant alors potentiellement des réductions temporelles. Pour ces dernières on peut chercher des explications, la première articulatoire : le phonème, facile et rapide à réaliser, a une durée intrinsèque courte. Un débit rapide avec une articulation incomplète réduit alors encore cette durée. Une deuxième explication peut venir des fréquences d'occurrence : une observation très fréquente est une observation à contenu d'information faible et risque donc d'être négligée dans le signal acoustique. Il est fort plausible que dans une parole spontanée ces deux facteurs se trouvent combinés.

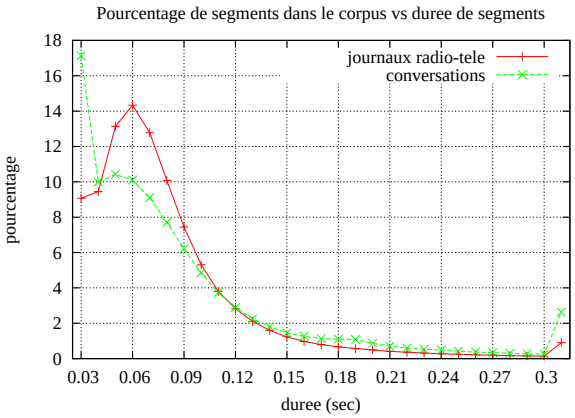


FIG. 6: Distribution de la durée des segments phonémiques pour deux styles de parole : parole préparée d'émissions journalistique et parole spontanée de conversations téléphoniques.

La figure 6 montre clairement l'effet du style de corpus. Pour le corpus journalistique la distribution de durée des segments réalise un pic globalisant 14% du corpus à 60 ms, alors que pour la parole conversationnelle ce pic se trouve à 50 ms avec seulement 10% des données. Mais pour ce dernier le maximum des durées est concentré sur

les durées courtes avec 17% des segments sur la durée minimale de 30 ms imposée par l'alignement automatique (modèles de Markov cachés à trois états avec transition entre états correspondant à 10 ms). Pour la parole journalistique ce taux est « seulement » de 9%. Tous les phonèmes ne contribuent pas tous de manière égale. La table 7 montre les phonèmes réalisant les plus forts pourcentages de durée minimale dans le corpus de conversations téléphoniques. Des taux plus faibles sont observés pour le corpus journalistiques, mais l'ordre des phonèmes impliqués reste sensiblement identique. Concernant les consonnes on trouve en premier lieu le /l/ avec 43% des segments à durée minimale. Le /l/ est très fréquent en français, en particulier sur les mots outils [4]. De manière générale les liquides, les semi-voyelles, le /v/, le /d/ et le /n/ sont les consonnes les plus affectées. Le /d/ comme le /l/ sert fréquemment de support aux mots outils. Pour les voyelles on voit que le schwa arrive en première position avec pratiquement 50% des occurrences. Ceci confirme par mesures détournées la nature instable du schwa. Ce qui est plus intrigant c'est de voir des pourcentages élevés pour les voyelles ouvertes /ɛ/ et /a/. Concernant le /a/ nous pouvons observer par exemple que ce segment peut quasiment disparaître dans des mots ou séquences de mot comme `voilà, parce que` réalisés plutôt comme `v'là, p'ce que`. De même les voyelles antérieures fermées sont alignées fréquemment avec une durée minimale. On peut remarquer que bon nombre de phonèmes qui ont parmi les plus forts pourcentages de durée minimale, sont également parmi les plus fréquents. Il manque cependant les consonnes non-voisées comme /t/ et /s/ qui ont une durée intrinsèque plus longue.

TAB. 7: Phonèmes réalisant les plus forts pourcentages de durée minimale dans le corpus de conversations téléphoniques. Ces pourcentages sont donnés pour les voix d'hommes et les voix de femmes.

cons	*% durées min.*		*voy*	*% durées min.*	
phon	*hom*	*fem*	*phon*	*hom*	*fem*
l	43	44	ə	50	46
ɥ	34	43	ɛ	30	24
v	33	26	a	26	16
j	28	21	i	25	20
R	26	23	y	24	27
d	23	17	e	24	16
n	20	12	ø	22	21

Cette première analyse des durées met en évidence un phénomène de réduction temporelle en parole spontanée et en donne une première quantification. Ce phénomène qui est certes connu, nécessite cependant dans le futur des investigations plus approfondies en lien avec des études sur les variantes phonologiques du français, la prosodie, le débit et la modélisation des prononciations pour le traitement automatique. En particulier des études en fonction des contextes phonémiques et des mots supports permettront d'éclairer si la réduction temporelle est plutôt conditionnée par les contextes phonémiques ou par la fréquence lexicale.

3.5. Prosodie

Nous terminons par un travail engagé dans le contexte du projet CNRS TCAN Varcom et du projet ANR PFC-Cor sur les variétés régionales du français. Le corpus PFC vise à proposer des échantillons représentatifs des parlers normatifs et vernaculaires d'un grand nombre de variétés de l'espace francophone via des enregistrements de parole lue et d'entretiens libres et guidés. Nous proposons ici une étude montrant comment à partir de grands corpus et de traitement automatique on peut partir à la recherche d'indices prosodiques caractérisant les accents régionaux.

L'accent suisse, représenté par le canton de Vaud (autour de la ville de Nyon), a pu être identifié facilement face à des parlers français méridionaux (Aix-Marseille, Douzens, Biarritz) et du nord (Vendée, Brécey) lors de tests perceptifs [31] autant par des sujets français de la région parisienne, que par ceux de la région d'Aix-Marseille. Est-ce que cette facilité d'identification pourrait être liée à des aspects prosodiques ? Une étude de la distribution des durées de segments (pris globalement, mais aussi en distinguant voyelles et consonnes) ne montre pas de différence significative pour le pays de Vaud. Or, d'après différentes études (p.ex. [23]) les schémas intonatifs du canton de Vaud semblent présenter quelques particularités, en particulier une tendance à l'accentuation de la première syllabe d'un mot bisyllabique, qui entraîne en général une montée de la courbe mélodique. D'après ces études cette tendance peut être liée au substrat franco-provencal, caractérisé par une accentuation de la pénultième syllabe. Eventuellement, sur un plan diatopique, ceci peut trahir un contact avec la langue allemande.

Le corpus du texte lu, segmenté à la fois en mots et en phonèmes permet d'étudier l'évolution de la courbe de F0 aux frontières de phonèmes et de mots. Un étiquetage en parties du discours permet d'affiner les analyses. A partir de la segmentation phonémique automatique de douze points du corpus PFC (notamment Aveyron-Paris [5], Biarritz, Brécey, Brunoy, Dijon, Douzens, Lacaune, Lyon-Villeurbanne, Nyon, Roanne, Rodez et Vendée), nous avons pu mesurer à l'aide de PRAAT [29] la fréquence fondamentale moyenne par segment. Dans l'idée d'étudier de manière plus détaillée le bigramme `déterminant nom`, nous avons ensuite associé des parties du discours aux mots du texte lu. La table 8 montre le nombre de bigrammes extraits par région, pour lesquels la fréquence fondamentale est définie à la fois sur la voyelle du déterminant et sur la voyelle de la première syllabe du nom.

TAB. 8: Nombre de bigrammes `déterminant nom` extraits du texte PFC.

Point d'enquête	*#bigrammes Det Nom*
Aveyron-Paris	155
Biarritz	186
Brécey	188
Brunoy	188
Dijon	137
Douzens	194
Lacaune	194
Lyon-Villeurbanne	193
Nyon	233
Roanne	161
Rodez	105
Vendée	93

La figure 7 montre la distribution des séquences *dét-nom* en fonction de la différence de F0 (ΔF_0) mesurée entre la première voyelle du nom et la voyelle du déterminant (qui précède immédiatement). Les mesures sont obtenues à partir du sous-corpus de PFC correspondant au texte lu qui est le même pour tous les locuteurs. Pour cette étude les voix d'hommes seules sont exploitées. La courbe en gras correspond à la distribution moyenne obtenue à partir

[4] Ceci peut être mis en relation avec des erreurs observées dans la première partie (p.ex. la contraction de `c'est les` en `ses`).

[5] Le point d'enquête Aveyron-Paris regroupe des locuteurs aveyronnais vivant depuis de nombreuses années à Paris.

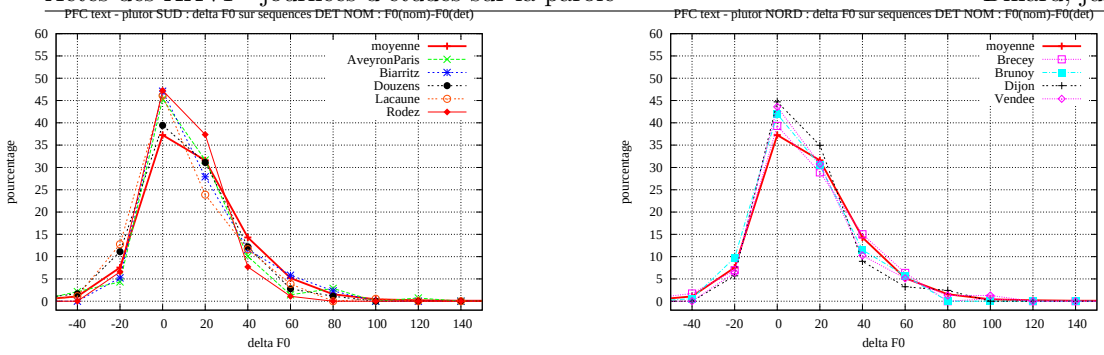


FIG. 7: ΔF_0 sur les séquences de partie de discours **déterminant nom** à partir du texte lu du corpus PFC ; **à gauche :** points d'enquête localisés plutôt au sud (Aveyron-Paris, Biarritz, Douzens, Lacaune, Rodez) ; **à droite :** points d'enquête localisés plutôt au nord (Brécey, Brunoy, Dijon, Vendée).

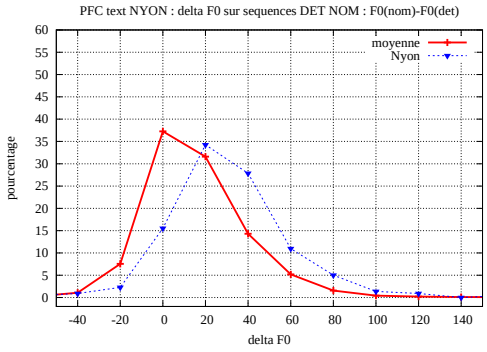


FIG. 8: ΔF_0 sur les séquences de partie de discours **déterminant nom** à partir du texte lu du corpus PFC : courbe moyenne et Nyon (pays de Vaud).

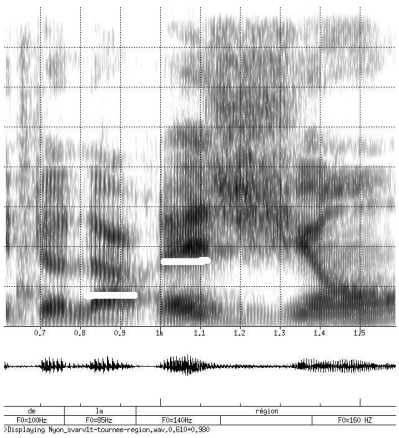


FIG. 9: Spectrogramme illustrant un ΔF_0 positif sur la séquence *dét-nom* : `la région` du locuteur `svarv` du pays de Vaud, extraite du corpus PFC.

de 12 points d'enquête. A gauche on voit les points d'enquête du sud et à droite plutôt du nord de la France. Les courbes ne varient pas significativement en fonction d'une séparation nord/sud. Les courbes montrent que sur les séquences *dét-nom* en français standard, ΔF_0 est centrée autour de 0 et reste proche de 0. Ainsi presque 40% des séquences *dét-nom* ont un ΔF_0 de zéro et pour 30% des données ΔF_0 vaut 20Hz. Seulement 15% des séquences *dét-nom* ont un ΔF_0 de 40Hz. La figure 8 compare la courbe moyenne au canton de Vaud. Le français du pays de Vaud se distingue dans cette étude par une montée de F0 qui commence dès la première syllabe du nom : 35% des 233 séquences examinées ont un ΔF_0 de 20 Hz et presque 30% atteignent 40Hz.

Ce premier résultat semble indiquer une spécificité prosodique qui irait dans le sens d'une accentuation de la première syllabe (ou amorce d'accentuation dès la première syllabe). Des études sur plus de données et avec des patrons plus complexes qu'un simple bigramme de partie de discours sont prévues, en particulier pour l'étude de la pénultième syllabe du nom.

4. CONCLUSION

Dans cette contribution nous avons essayé de donner un aperçu de quelques problématiques de recherche en transcription automatique et en analyse de grands corpus oraux de différents genres, avec des perspectives de recherche à la fois pour le traitement automatique et des domaines plus linguistiques. Nous avons montré une analyse des erreurs de transcription sur des données journalistiques, une étude similaire pour la parole conversationnelle reste à faire. L'importance du modèle de langage dans bon nombre de confusions d'homophones a été mis en évidence notamment par une mesure modifiée du taux d'erreur de mot. Les ambiguïtés dues aux homophones expliquent une large part des erreurs en parole journalistique. Une partie des erreurs peut cependant être attribuée à une modélisation acoustique inadéquate des mots. En particulier des contractions temporelles mettent en défaut la modélisation des mots comme séquences de phonèmes d'une prononciation standard. Alors que ce problème se pose en termes relativement discrets pour la parole journalistique, il devient flagrant sur la parole conversationnelle. Cette dernière est à débit plus variable, parfois très rapide, parfois très lente, ce qui est illustré par une distribution des durées phonématiques globalement plus étalée, avec une concentration élevée de segments sur la durée minimale de 30 ms. A cet endroit la courbe cumule tous les segments attendus par le modèle de prononciation et qui n'ont pas (ou peu) été réalisés dans la parole effectivement produite. Des résultats d'analyses de corpus présentées ici concernent les fréquences d'occurrences des phonèmes dans les corpus et nous avons pu montrer que ces fréquences varient en fonction du genre de corpus traité et également en fonction du contexte phonémique gauche et droit. Ainsi pour le corpus de données journalistiques nous avons présenté les comptes d'occurrences des voyelles en contextes gauche et droit labial, alvéolaire, postalvéolaire, palato-vélaire et des liquides /l/ et /R/.

Les progrès accomplis en traitement automatique permettent d'aborder bon nombre de recherches sous un angle nouveau. La disponibilité de corpus et d'instruments pour

l'accès au contenu permet de poser un nombre élevé de questions en même temps et d'avoir très vite, si ce n'est une réponse, au moins une tendance. Nous vivons actuellement une révolution technologique qui permettra d'enrichir le domaine de la linguistique de l'oral de nouveaux instruments et de méthodologies expérimentales exploitant de grands corpus [24]. L'ère chomskyenne a rendu pendant des décennies l'usage de corpus en linguistique pour le moins suspect, si ce n'est hors sujet. Sans vouloir rentrer dans des polémiques scientifiques, force est de constater que nous sommes aujourd'hui à un tel point d'accès facile à des données orales qu'il serait non scientifique de refuser l'étude de ces données, dont le corpus ESTER est certainement un exemple important pour le français. De telles études nous pouvons espérer dégager de nouvelles connaissances sur la langue orale et les performances des locuteurs en lien avec la neuro- et psycholinguistique. Ces connaissances seront à terme certainement utiles pour les systèmes de traitement automatique de la parole au sens large, incluant au-delà de la transcription des problématiques comme l'identification des locuteurs, des langues et des accents, la synthèse, la compréhension et le dialogue.

REMERCIEMENTS

Partie des travaux ont été réalisés grâce aux projets interministériel TECHNOLANGUE-ESTER, CNRS TCAN Varcom, ANR PFC-Cor et au projet européen CHIL. Je tiens à remercier ici mes collègues du LIMSI, de la DGA et de Paris 3 qui ont contribué par leurs travaux, réflexions, critiques et suggestions aux résultats présentés.

RÉFÉRENCES

[1] M. Adda-Decker, L. Lamel, "*Pronunciation Variants Across System Configuration, Language and Speaking Style*", Speech Communication "Special Issue on Pronunciation Variation Modeling", **29**, 1999.

[2] M. Adda-Decker, P. Boula de Mareüil, L. Lamel, "*Pronunciation variants in French : schwa & liaison*", 14th International Conference on Phonetic Science, ICPhS-99, août 1999.

[3] M. Adda-Decker, Ph. Boula de Mareüil, G. Adda, & L. Lamel. "Investigating syllabic structures and their variation in spontaneous French", *Speech Communication*, 46 (2005) pp.119-139, Elsevier ed.

[4] M. Adda-Decker, L. Lamel, "Do Speech Recognizers Prefer Female Speakers ?", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[5] C. Barras et al., "Transcriber : development and use of a tool for assisting speech corpora production". Speech Communication, 33(1-2) :5-22, Jan 2001.

[6] C. Blanche-Benveniste, "Constitution et exploitation d'un grand corpus", Rev. Française de linguistique appliquée, 1999, IV-1 (65-74).

[7] L-J. Boë, J-P. Tubach, "Une base de données lexicale orthographique-phonétique du français parlé". Cahiers de grammaire 17, Universite de Toulouse-Le Mirail, novembre 1992.

[8] P. Boula de Mareüil et al., "Liaisons in French : a corpus-based study using morpho-syntactic information". In Proceedings of the International Conference on Phonetic Sciences, ICPhS, Barcelone août 2003.

[9] P. Boula de Mareüil et al. "A quantitative study of disfluencies in French broadcast interviews", dans Proceedings DISS'05, Aix-en-Provence, septembre 2005.

[10] M. Candea et al., "Inter- and intra-language acoustic analysis of autonomous fillers", dans Proceedings DISS'05, Aix-en-Provence, septembre 2005.

[11] P. Delattre (1966) Studies in French and Comparative Phonetics, La Haye, Mouton.

[12] F. Dell (1973). Les règles et les sons : introduction à la phonologie générative. Paris : Hermann. 2e éd.

[13] J.M. Dolmazon et al., *"Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale"*, JST97, Avignon, avril 1997.

[14] D. Duez, 2003. Modelling Aspects of Reduction and Assimilation in Spontaneous French Speech, In Proc. IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition, 2003. Tokyo.

[15] J. Durand, B. Laks, C. Lyche, (2003). Le projet « Phonologie du français contemporain » (PFC). La Tribune Internationale des Langues Vivantes 33 3-9.

[16] P. Encrevé, (1988). La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français. Éditions du Seuil, Paris.

[17] P. Fouché, (1959). Traité de prononciation française. Editions Klincksieck, Paris.

[18] C. Fougeron et al.,"Liaison and schwa deletion in French : an effect of lexical frequency and competition", *Eurospeech*, Aalborg (pp. 639-642), 2001.

[19] S. Galliano et al. "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[20] J.L. Gauvain et al., "Speaker-Independent Continuous Speech Dictation," Speech Communication, **15** :21-37, Sept. 1994.

[21] J-L. Gauvain et al., "Where Are We In Transcribing French Broadcast News ?", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[22] C. Gendrot, M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in French and German", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[23] M-A. Hintze, T. Pooley & A. Judge (eds.). "French accents : Phonological and sociolinguistic perspectives", pub. CILT/AFLS, ISBN 1 909031 95 4, 2001.

[24] B. Habert, "Portrait de linguiste(s) à l'instrument", Texto ! Textes et cultures, ISSN 1773-0120, Vol. X, n.4, décembre 2005.

[25] L.F. Lamel et al., "BREF, a Large Vocabulary Spoken Corpus for French," *EuroSpeech'91*.

[26] L.F. Lamel et al. "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech-95*, Madrid, septembre 1995.

[27] Malécot A. (1974) Frequency of occurence of French phonemes and consonant clusters, Phonetica 29.

[28] N. Nguyen et al., "Detection of liaison consonants in speech processing in French : Experimental data and theoretical implications", Laboratory Approaches to Romance Phonology, édité par P. Prieto et M.J. Solé (John B Benjamins), à paraître.

[29] PRAAT, a system for doing phonetics by computer. Glot International 5(9/10) : 341-345, 2001

[30] I. Vasilescu et al., "Hésitations autonomes dans 8 langues : une étude acoustique et perceptive", Colloque MIDL04 Paris, 29-30 novembre 2004.

[31] C. Woehrling, P. Boula de Mareüil , "Perceptual identification of French varieties in the PFC corpus", 7e Rencontres Internationales du Réseau Français de Phonologie (RFP), Aix-en-Provence, 2005.

[32] S.J. Young et al., "*Multilingual large vocabulary speech recognition : the European SQALE project*", Computer Speech & Language, vol. 11, nb.1, janv. 1997.

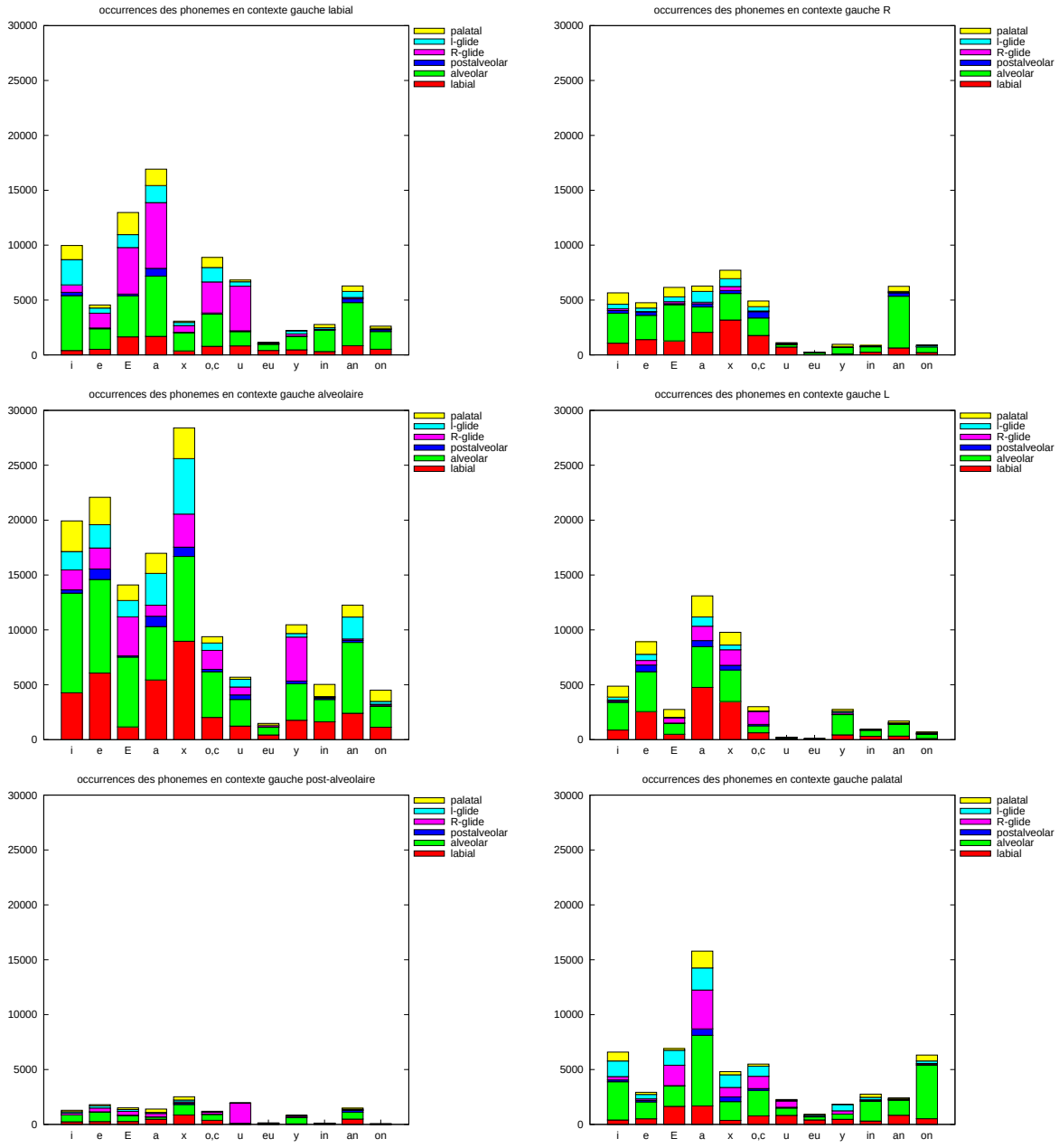


FIG. 10: Fréquences d'occurrence des voyelles en contextes gauche et droit de labiales, d'alvéolaires, de postalvéolaires, de la liquide /R/, de la liquide /l/ et de palato-vélaires mesurées sur le corpus **journalistique**. **En haut à gauche :** le contexte gauche correspond aux **labiales** ; **à droite :** le contexte gauche correspond à **/R/** ; **milieu à gauche :** le contexte gauche correspond aux **alvéolaires** ; **à droite :** le contexte gauche correspond à **/l/** ; **en bas à gauche :** le contexte gauche correspond aux **postalvéolaires** ; **à droite :** le contexte gauche correspond aux **palato-vélaires**. Pour chaque contexte gauche les histogrammes indiquent la distribution des voyelles en fonction des 6 classes de consonnes en contexte droit.

Session XV

Corpus et variabilité

Jeudi 15 juin 2006 - 10h00 11h00

Détection automatique de frontières prosodiques dans la parole spontanée

Katarina Bartkova, Natalia Segal

France Télécom R&D/TECH/SSTP
2 av. Pierre Marzin, 22307 Lannion Cedex, France
katarina.bartkova@francetelecom.com, natalia.segal@francetelecom.com

ABSTRACT

The present study addresses the issue of the automatic detection of prosodic units in French. An analysis of two prosodic parameters, phone duration and F0 slope values, carried out on two spontaneous speech databases recorded by several thousands of speakers, revealed relevant deviations of these parameters at prosodic junctions. Vowel durations and F0 slopes are used to automatically detect prosodic units at the two data bases. The phone duration is modelled as the ratio of two subsequent vowel durations. Apart from the duration ratio, duration values are also modelled. The detection of prosodic units based on the F0 uses the value of the F0 slope and its standard deviation, recalculated after each pause. An evaluation of the automatic detection is carried out by comparing the prosodic border locations with the lexical boundaries and also with prosodic boundaries obtained by manual segmentation.

1. INTRODUCTION

La détection automatique des événements prosodiques à partir du signal acoustique constitue un pas important vers la construction d'une relation univoque entre le signal acoustique et la représentation abstraite de la prosodie. Plusieurs approches ont été proposées pour construire une telle relation dont l'importance serait primordiale pour toute application en traitement automatique de la parole. Certaines de ces approches sont basées sur des principes perceptifs [6], d'autres sur des principes articulatoires [3]. Toutefois, la plupart des modèles sont entrainés à partir d'un corpus plus ou moins important selon une méthode statistico-probabiliste [4,5].

Des tentatives d'utilisation des paramètres prosodiques en reconnaissance de la parole ont été entreprises avec plus ou moins de succès. Certaines études visaient la détection des unités prosodiques pour réduire l'espace de recherche des candidats lexicaux lors du décodage phonétique du signal de parole [7]. Le but recherché par cette démarche est la diminution de la perplexité de la tâche qui entraîne en général une amélioration des performances. L'utilisation des paramètres prosodiques a également été testée lors du post-traitement pour confirmer ou infirmer les hypothèses de reconnaissance proposées par le décodeur [2]. Même si certains systèmes montrent un progrès considérable, la construction d'un codage prosodique, général et suffisamment fiable pour les applications automatiques, reste à réaliser. Cette tâche est encore plus ardue pour la parole spontanée dont la prosodie est extrêmement variable.

Nous présentons dans cette étude deux techniques de détection des frontières des unités prosodiques à partir du signal acoustique de la parole, destinées à une utilisation en reconnaissance de la parole. L'une des méthodes est basée sur l'utilisation de la durée phonémique et l'autre sur l'évolution de la fréquence fondamentale.

2. CORPUS UTILISÉS

Dans cette étude, nous avons utilisé deux bases de données de la parole continue spontanée enregistrées à travers le réseau téléphonique. Le premier corpus est un corpus d'Enquêtes de Satisfaction (corpus ES) et le second un corpus de Messages Courts (corpus MC) destinés à être transmis sous forme de texte (SMS). Les deux corpus sont constitués de monologues de longueurs variables. Le premier corpus contient environ 1000 enregistrements (~50 mots par message) et le deuxième environ 9000 enregistrements (~18 mots par message). On suppose que chaque enregistrement est prononcé par un nouveau locuteur.

Les deux corpus ont été manuellement retranscrits. La forme phonétique des textes a été par la suite alignée automatiquement avec le signal de parole simulant ainsi la sortie d'une reconnaissance "parfaite", sans erreurs. Cet alignement nous a permis d'accéder aux paramètres prosodiques des phonèmes.

3. ANALYSE STATISTIQUE PRÉLIMINAIRE

L'étude statistique a été focalisée sur la durée syllabique et vocalique ainsi que sur les variations de la courbe de F0, car ces paramètres jouent un rôle primordial dans la perception de la structuration prosodique.

3.1. Analyse de la durée phonémique

Afin d'analyser le comportement de la durée phonémique, nous avons comparé les distributions de la durée normalisée des noyaux syllabiques selon leur position dans le mot et dans des unités prosodiques. Nous considérons ici comme unité prosodique la portion de signal de parole délimitée par deux pauses. Cette simplification s'est avérée pratique pour le traitement de la parole spontanée où le nombre des pauses est relativement élevé. La comparaison des distributions avec le test ANOVA (***AN***alysis ***O***f ***VA***riance) a démontré, pour les deux corpus, un allongement très significatif ($p < 0,001$)

et très important de la durée à la fin de l'unité prosodique et un allongement un peu moins important mais toujours significatif à la fin du mot.

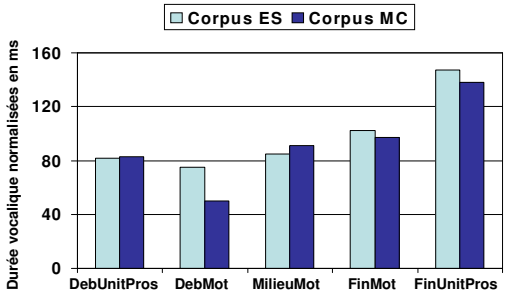


Figure 1 : Distribution de la durée vocalique selon la position du noyau vocalique.

L'interprétation possible de ces résultats est que les syllabes avant les pauses correspondent le plus souvent à une frontière prosodique, par conséquent leur durée s'allonge. Or, les syllabes finales des mots, non suivies de pause, ne coïncident pas toujours avec une frontière prosodique (d'où une plus grande dispersion de leurs durées).

3.2. Analyse de la fréquence fondamentale

Les distributions de la pente de F0 sont également comparées sur les syllabes selon leur position. Le mouvement de F0 a été représenté par sa direction, sa pente (la vitesse de changement) et son amplitude absolue mesurée en semi-tons [7]. L'amplitude absolue du mouvement s'est révélée être le paramètre le plus pertinent. Les résultats du test ANOVA, pour les amplitudes du mouvement de F0, ont montré que la valeur moyenne en fin d'unité prosodique est très significativement ($p < 0{,}001$) plus grande que dans toutes les autres positions et cela pour les deux corpus.

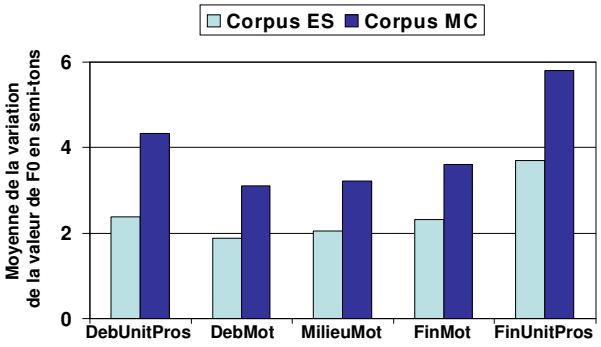


Figure 2 : Distribution de la variation de la valeur de F0 (amplitude absolue) selon la position de la syllabe.

Comme pour la durée vocalique, ici encore, nous pouvons émettre l'hypothèse que les mouvements importants de F0 marquent les frontières prosodiques : c'est presque toujours le cas avant une pause (sauf en cas d'hésitation [1]), et c'est également vrai pour certains mots à l'intérieur d'une unité prosodique.

4. DÉTECTION DES FRONTIÈRES PROSODIQUES

L'analyse statistique a démontré une spécificité de la valeur de la durée phonémique et du mouvement de F0 sur des frontières prosodiques. Pour capter cette spécificité nous avons entrepris la modélisation de ces paramètres afin de segmenter le signal de parole.

Les résultats de la segmentation en unités prosodiques sont évalués d'une part par rapport aux frontières lexicales et d'autre part par rapport aux frontières prosodiques annotées manuellement sur un sous-ensemble (20%) de données du corpus ES. La décision de tester le découpage automatique par rapport aux frontières lexicales a été motivée par le souci de relever des erreurs de découpage sur une base de données de taille conséquente. Dans ce test-là il ne s'agit pas d'examiner si l'emplacement d'une frontière prosodique est correct, mais de vérifier que la frontière prosodique hypothétique (proposée par le découpage automatique), coïncide avec la frontière du mot ou non. Si elle coïncide avec la frontière lexicale elle est acceptée comme potentiellement correcte, et si elle ne coïncide pas avec la frontière lexicale (dernière voyelle du mot), elle est comptabilisée comme erreur de découpage. Bien que cette dernière évaluation puisse être discutable, il faut remarquer qu'une segmentation manuelle effectuée par plusieurs experts ne fournit pas nécessairement les mêmes résultats. En effet dans beaucoup de cas la détection des frontières prosodiques dépend de l'appréciation subjective de l'expert qui segmente [5].

Lors de l'évaluation de l'emplacement des frontières prosodiques nous nous focaliserons sur les frontières non suivies de pauses; les frontières prosodiques suivies de pause sont exclues de l'analyse.

4.1. Durée phonémique

Une modélisation discrète de la durée phonémique a été réalisée pour deux positions pertinentes: position frontière prosodique et position non-frontière prosodique. Afin d'éviter de segmenter manuellement en unités prosodiques une base de données importante, nous avons utilisé pour l'apprentissage de nos modèles deux positions facilement repérables automatiquement et qui correspondaient néanmoins d'une façon relativement exacte à ces deux positions. Nous avons utilisée la position interne des mots pluri-syllabiques pour l'apprentissage du modèle "non-frontière-prosodique" et la position "fin de mot suivie d'une pause" pour l'apprentissage du modèle "frontière-prosodique"

La modélisation de la durée phonémique a été réalisée uniquement pour les voyelles. Cette démarche se justifie par le fait que les durées vocaliques sont plus facilement comparables entre elles que les durées syllabiques, car la syllabe possède une structure de complexité variable ayant un impact direct sur sa durée. La modélisation de la durée

a été réalisée à travers une modélisation discrète par la construction d'histogrammes normalisés des trois paramètres suivants : durée de la voyelle se trouvant sur la frontière prosodique hypothétique, durée de la voyelle qui suit l'hypothèse de frontière et le rapport de ces deux durées vocaliques. Le rapport des durées des deux voyelles adjacentes (voyelle courante et voyelle suivante) s'est avéré une méthode simple à implémenter et efficace dans la recherche des frontières prosodiques.

Utilisation et évaluation du modèle

Au moment du test, le critère de décision de l'occurrence d'une frontière prosodique (FP) basé sur un rapport de vraisemblance "**P(param|FP)/P(param|non_FP)" ,** supérieur à un seuil prédéfini, et complété par un test sur le rapport des durées (pour une frontière prosodique, le rapport doit être supérieur à 1). Ainsi, pour qu'une frontière prosodique soit détectée par cette méthode, deux conditions doivent être réunies : le rapport des durées doit être plus grand que 1 (condition requise pour l'occurrence d'une frontière) et le score fourni par le modèle doit être plus élevé que le seuil de décision arbitraire utilisé. Si le rapport de vraisemblance est inférieur à ce seuil, la possibilité de l'occurrence d'une frontière prosodique est rejetée, sinon elle est acceptée.

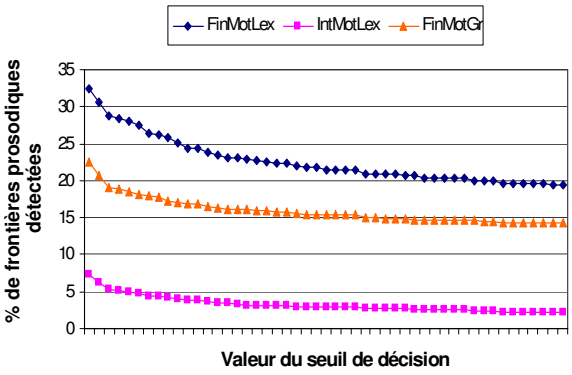


Figure 4 : Evolution des hypothèses de frontières prosodiques en fonction de la valeur du seuil choisi.

Lors des évaluations préliminaires, des tests d'hypothèses ont été appliqués pour différentes valeurs du seuil de décision, d'une part sur les frontières des mots et d'autre part sur des positions internes aux mots. Une différence est faite entre mots grammaticaux (clitiques monosyllabiques) et mots lexicaux. La figure 4 trace les résultats obtenus. Selon la valeur du seuil, 20 à 30% des fins des mots lexicaux sont détectées comme frontières prosodiques possibles. En revanche, pour les positions internes aux mots (IntMotLex), qui ne peuvent pas être des frontières prosodiques et constituent donc des erreurs incontestables de détection, seules 3 à 5% d'entre elles sont détectées à tort comme des frontières prosodiques.

Ces mesures de taux de détection d'hypothèses de frontières prosodiques ont également été réalisées par des tests croisés entre les deux bases de données utilisées. L'apprentissage des paramètres a été effectué sur la moitié des données appartenant soit à la même base que les données de test, soit à l'autre base. Comme cela apparaît au vu des résultats de la Table I, le seuil de décision des frontières prosodiques est dépendant de la base d'apprentissage. Néanmoins, si nous options pour un seuil de décision fournissant un taux d'erreur comparable de **4%,** le nombre de frontières prosodiques hypothétiques, détectées sur le corpus **MC** serait alors de **26%** pour les mots lexicaux et **10%** pour les mots grammaticaux.

Table I: Détection des frontières prosodiques avec bases d'apprentissage et de tests croisées.

Corpus		Fin Mot Lexical	Fin Mot Gram	InterMot (Erreur)
Tst	**App**			
ES	ES	26,3%	17,8%	***4,3%***
	MC	17,1%	12,9%	***1,9%***
MC	ES	30,5%	10,7%	***4,0%***
	MC	19,9%	5,8%	***1,7%***

4.2. Fréquence fondamentale

L'évolution de la courbe de la fréquence fondamentale a été employée pour découper le signal de parole en unités prosodiques. Dans cette approche nous n'avons pas utilisé de modélisation proprement dite, ainsi elle est exempte d'apprentissage. La décision de détection d'une frontière prosodique est prise en fonction de la pente de la fréquence fondamentale observée et de l'écart-type de la variation de la valeur de la fréquence fondamentale sur la portion de signal de parole située entre deux pauses. Dans cette approche uniquement les pentes montantes sont considérées. La valeur de l'écart-type de F0 est remise à jour à chaque nouvelle pause. Le seul préalable pour l'utilisation de cette approche est la détection des pauses présentes sur le signal de parole (détection bruit/parole). Le seuil de détection des frontières prosodiques est exprimé en % de l'écart-type.

Evaluation de la méthode

Comme la technique du découpage prosodique par la pente de F0 n'utilise pas le décodage phonétique du signal de parole et exploite tous les segments voisés, certains critères d'évaluation sont différents. Le découpage prosodique obtenu par la pente de F0 est considéré correct quand il se situe sur le nucleus ou la coda de la dernière syllabe d'un mot – c'est-à-dire sur la fin du mot. Tous les autres emplacements sont considérés comme erronés. Quand la frontière prosodique est placée à tort alors l'erreur moyenne de détection est quantifiée en ms comme l'écart entre la frontière du mot et la frontière prosodique.

La Table II indique, pour les deux bases de données et pour un fonctionnement donné du seuil de décision, le taux de frontières prosodiques non-suivies de pauses et pour ces frontières-là le taux dont l'emplacement coïncidait avec une frontière lexicale. La dernière colonne comporte l'écart moyen d'erreur entre la frontière détectée et la frontière lexicale en ms (dont l'écart-type se situe à ~

70 – 80 ms).

Table II: Détection des frontières prosodiques (FP) par la pente de F0.

Corpus	FP non suivi de pause	FP-Frontière Lexical	Erreur
ES	43%	80%	30 ms
MC	23%	77%	36 ms

4.3. Comparaison à la segmentation manuelle

Afin d'estimer la précision du découpage prosodique automatique, 20% du corpus ES a été segmenté manuellement en unités prosodiques. 48% des frontières prosodiques, placées manuellement, étaient des frontières non-suivies de pauses. La comparaison entre segmentation manuelle et automatique est effectuée uniquement pour les frontières prosodiques non suivies de pause. Afin d'effectuer cette comparaison, le point de fonctionnement des méthodes automatiques a été choisi de sorte qu'il représente un bon compromis entre le nombre de frontières prosodiques placées à tort et le nombre de frontières prosodiques placées sur des frontières lexicales.

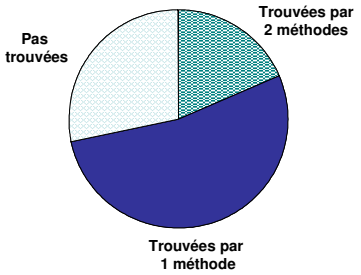


Figure 5 : Evaluation de la détection des frontières prosodiques en fonction des frontières manuelles.

Pour le point de fonctionnement choisi, la segmentation automatique utilisant la durée phonémique a abouti à un nombre de frontières prosodiques non suivies de pause (1242) inférieur à la segmentation manuelle (1446) alors que la segmentation par F0 a abouti à un nombre supérieur (2001). Une correspondance entre découpage manuel et découpage automatique obtenu par la durée phonémique concernait 44% des unités prosodiques (non suivies de pause), et entre découpage manuel et découpage par F0 concernait 46% des unités prosodiques (non suivies de pause). La répartition de la complémentarité et de la correspondance entre les 3 méthodes de détection des frontières prosodiques est illustrée sur la Figure 5. Ainsi, 18,5% des frontières prosodiques manuelles correspondaient avec les frontières prosodiques détectées par les deux paramètres, tandis que 53% des frontières prosodiques placées manuellement correspondaient avec une frontière prosodique détectée par une des deux méthodes de découpage automatique. Cela démontre qu'une complémentarité entre les deux paramètres existe quant à la détection des frontières prosodiques. Ainsi, par exemple, 80% des hésitations présentes sur le signal de parole ont été détectées comme frontière prosodique par la durée vocalique. Or, une frontière prosodique a été placée sur moins de 10% des hésitations (8 % pour la base ES et 6,7% pour la base MC) quand la détection de la frontière prosodique a été réalisé par la pente de F0.

5. CONCLUSION

Nous avons présenté dans cette étude une analyse prosodique de deux bases de données de parole spontanée enregistrées par plusieurs milliers de locuteurs. Une méthode de découpage automatique du signal de parole en unités prosodiques a été développée en utilisant la durée vocalique et la pente de F0. L'évaluation des résultats de cette méthode a été effectuée par rapport aux frontières lexicales et aux frontières prosodiques placées manuellement. Le découpage automatique a donné des résultats plus qu'encourageants. Par ailleurs, une complémentarité des paramètres dans la détection des frontières prosodiques a été observée.

La suite de ce travail devrait s'orienter vers l'étude d'une combinaison des deux paramètres tendant vers leur utilisation conjointe dans la détection des frontières prosodiques. Par ailleurs, la fiabilité des paramètres devrait être renforcée par leur apprentissage sur une base segmentée manuellement en unités prosodiques.

BIBLIOGRAPHIE

[1] K. Bartkova. Prosodic cues of spontaneous speech in French. Dans *DISS'05,* pages 21-25, 2006.

[2] K. Bartkova, D. Jouvet. Usefulness of phonetic parameters in a rejection procedure of an HMM-based speech recognition system. Dans *EUROSPEECH-1997*, pages 267-270, 1997.

[3] H. Fujisaki. Information, Prosody, and modeling with emphasis on tonal features of Speech. Dans *Speech Prosody 2004,* pages 1–10, 2004.

[4] D. Hirst, A. Di Cristo and R. Espesser. Levels of representation and levels of analysis for the description of intonation systems. Dans *Prosody: Theory and Experiment.* Kluwer Academic Press, Dordrecht, 2000.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg. TOBI: A standard for labeling English prosody. Dans *Proceedings ICSLP 92*, pages 867-870, 1992

[6] J. 't Hart, R. Collier, A. Cohen. *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge University Press, 1990.

[7] A. Waibel. *Prosody and speech recognition*, Morgan Kaufmann, London, 1988

Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique.

Cédric Gendrot et Martine Adda-Decker***

* Laboratoire de Phonétique et Phonologie – CNRS/UMR7018
ILPGA, 19, rue des bernardins – 75005 Paris
** LIMSI-CNRS bât. 508, BP 133, 91403 Orsay cedex
Mél: cgendrot@univ-paris3.fr - madda@limsi.fr

ABSTRACT

The aim of the present study is to highlight peripherality of French vowels in two prosodic positions: (i) word-final syllables (as compared to word-initial syllables), and (ii) in the vicinity of (before and after) pauses. The LIMSI speech alignment system is used [1] and formant values of oral vowels are automatically measured in a total of 25000 segments from two hours of journalistic broadcast speech in French. A tendency to reduction for all vowels (in terms of the shrinking of the vocalic triangle formed by F1 and F2 values) of short duration was clearly observed in a former study (Gendrot & Adda-Decker [2]). We show that at some extent, a similar relationship holds for vowels in both word-final and word-initial syllables.

1. INTRODUCTION ET METHODE

La disponibilité de grands corpus audio et d'outils automatiques d'alignement de plus en plus performants nous a permis de réaliser une étude (Gendrot et Adda-Decker [2]) sur les valeurs formantiques de 25000 voyelles orales en parole continue, et ce de manière automatique. Ces travaux contribuent à l'établissement de valeurs de formants et de leur variabilité en français, peu documentées à l'heure actuelle. Notre intérêt s'est porté principalement sur les variations de F1 et de F2 qui peuvent être interprétées – si on ne tient pas compte de l'effet des lèvres – en termes d'aperture/fermeture (corrélée à F1) et d'antériorité/postériorité (corrélée à F2). Nous avons montré :

- qu'il est possible d'extraire avec des traitements automatiques un pourcentage élevé (94%) des valeurs de formants de manière fiable.

- que l'espace vocalique formé par les formants F1 et F2 diminue progressivement avec la durée des segments analysés ; la durée étant répartie en trois catégories équilibrées (cf. Figure1). Ces variations sont plus faibles pour les voyelles fermées /i/ et /y/ qui sont, comme il est reconnu pour le français, mieux caractérisées par la proximité de F3 avec F2 (/y/) ou F4 (/i/).

- que les variations des formants étaient d'une amplitude comparable à celles observées pour des langues à accent lexical telles que l'allemand.

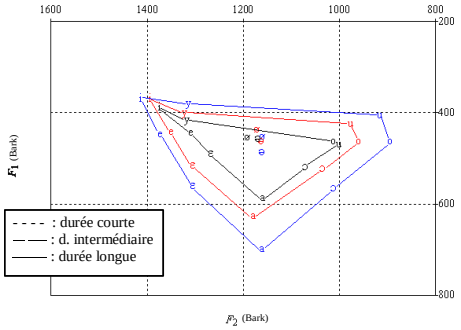


Figure 1 : Valeurs moyennes de F1 et F2 pour les voyelles orales du français en fonction de leur durée (normalisation en Bark).

Le corpus retenu est constitué d'enregistrements radio- et télédiffusés d'émissions journalistiques. Il s'agit de parole publique : l'articulation, sans être soutenue, y reste bonne, afin que la parole puisse être partagée par une large audience. La parole ne peut pas être qualifiée de spontanée, il s'agit plutôt de parole préparée : on observe peu d'hésitations, peu de fragments de mots et les structures syntaxiques restent souvent proches du langage écrit. Les phénomènes de réduction, auxquels nous nous intéressons à travers cette étude, y sont sans doute moindres que dans une vraie parole spontanée. Le corpus utilisé correspond à environ 2 heures de parole utile (15 hommes et 15 femmes) extraites pour la majeure partie d'émissions de France Inter, enregistrées en 1998 et transcrites orthographiquement (20k mots). Il s'agit ici de ressources utilisées dans le cadre du projet MIDL provenant du CTA/DGA, partenaire du projet.

Des analyses statistiques ont été effectuées au moyen d'ANOVAs à 1 facteur (position ou durée) voyelle par voyelle ; les tables de statistiques ne seront pas détaillées ici par manque de place.

2. HYPOTHÈSES

Cette étude vise à explorer plus précisément les variations formantiques des voyelles analysées en fonction de leur position prosodique. En effet, Gendrot et Adda-Decker [2] avaient conclu en comparant le français et l'allemand d'après un protocole expérimental identique à celui-ci, que des réductions vocaliques importantes pouvaient être mesurées en fonction de la durée à la fois en français et en allemand (cf. Figure 1). Le français n'étant pas une

langue à accent lexical, et typiquement considéré comme une langue « syllable-timed » avec un allongement final en fin de groupes de sens (ou groupes rythmiques), des phénomènes de réduction vocalique plus faibles pouvaient être attendus, tel que cela était suggéré par Delattre [3]. Nous analyserons séparément dans cette étude des positions prosodiques initiales et finales en français afin de vérifier si des variations de durée (réparties sur les trois catégories équilibrées de durée établies dans Gendrot & Adda-Decker [2]) impliquent également des variations formantiques, et de même ordre.

Deux facteurs pourraient expliquer un allongement vocalique en français : (i) La présence d'accents de focalisation/emphase qui, en français, se positionnent plus vraisemblablement sur les syllabes initiales de mots et qui impliquent – en plus d'une intensité accrue et d'une f0 augmentée – un allongement vocalique. (ii) La proximité de frontières est une cause fréquente d'allongements vocaliques et peut également être corrélée à la présence de pauses.

Dans ce présent travail, nous nous intéresserons au deuxième facteur de variation : dans un premier temps, les voyelles de syllabes finales de mots seront analysées puisque potentiellement plus longues, et comparées aux voyelles en syllabes initiales de mots (généralement non allongées, sauf en cas d'insistance). Nous séparerons également les valeurs en fonction des trois catégories de durée précédemment établies. Dans un deuxième temps, nous effectuerons les mêmes analyses formantiques du point de vue de la présence/absence de pauses dans l'entourage vocalique. Nous discuterons enfin de ces résultats en termes de prosodie articulatoire.

3. VARIATIONS FORMANTIQUES EN SYLLABE INITIALE/FINALE DE MOT

Nous avons ainsi séparé les mesures de formants en fonction de la position de la voyelle orale dans le mot en considérant les syllabes initiales (5200 occurrences) et finales (5300) de mots polysyllabiques. Pour ce faire, nous avons eu recours à la transcription à la fois orthographique et phonémique, en considérant comme règle de base que chaque voyelle alignée (les semi-voyelles étant donc exclues) correspond à une syllabe.

La figure 2 révèle que les voyelles en syllabes finales de mot ont des valeurs plus extrêmes que les syllabes initiales de mot sur les axes F1/F2, à l'exception des voyelles antérieures fermées /i/ et /e/. Les tables 1 et 2 détaillent la répartition des voyelles en fonction de leur position dans le mot, et en fonction de leur catégorie de durée.

Dans Gendrot [4], il a été suggéré que les procédures de normalisation telles que Lobanov, Nearey et Bark ne modifient sensiblement pas les résultats présentés ici de par une quantité importante et équilibrée de données. Une normalisation (Lobanov) sera toutefois utilisée pour faciliter la présentation des résultats cumulés hommes/femmes ici.

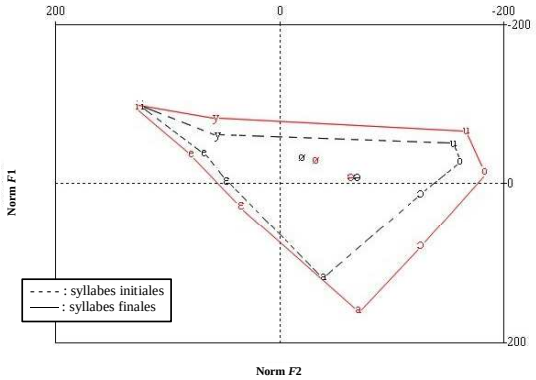


Figure 2 : Valeurs moyennes de F1 et F2 pour les voyelles en syllabe initiale et finale (Normalisation Lobanov, toutes durées confondues).

Afin de mieux interpréter ces résultats, nous avons également illustré une mise en regard de la distribution des voyelles en fonction de leurs valeurs de f0. Des catégories de f0 équilibrées (uniformes) ont ainsi été déterminées sur le même principe que les catégories de durée :

	f0 basse	f0 moyenne	f0 haute
hommes	<= 110 Hz	110 <f0 <= 140	> 140 Hz
femmes	<= 160 Hz	160 <f0 <= 210	> 210 Hz

Nous pouvons observer sur ces tables que les voyelles longues sont également celles qui ont des valeurs de f0 plus élevées, plus particulièrement en syllabe finale de mots[1]. Les voyelles longues sont également plus fréquentes lorsque positionnées en syllabe finale de mots.

Table 1 : répartition (en %) des voyelles en syllabe initiale de mots en fonction de leurs catégories de durée et de f0.

f0 / durée	bas	moyen	haut	total (durée %)
court	**14**	**14**	9	37
moyen	11	**14**	**14**	**39**
long	5	7	**12**	24
total(f0 %)	30	35	34	100

Table 2 : répartition (en %) des voyelles en syllabe finale de mots en fonction de leurs catégories de durée et de f0.

f0 / durée	bas	moyen	haut	total (durée %)
court	10	**11,5**	8,5	30
moyen	8	11	**14**	33
long	7,5	10	**19**	**37**
total(f0 %)	25	33	42	100

Les figures 3a et 3b (page 4) montrent les variations mesurées pour les voyelles en syllabes initiales et finales respectivement, tout en mettant en évidence les

[1] Notons que les intonations conclusives par nature plus basses ont été englobées dans ce résultat par manque d'étiquetage approprié, elles seront analysées dans une étude ultérieure.

différentes catégories de durées utilisées précédemment dans Gendrot et Adda-Decker [2], à savoir [30–50ms] pour les voyelles courtes, [60–80ms] pour les voyelles à durée intermédiaires et [90-110] pour les voyelles longues. Ces figures indiquent à première vue des tendances similaires, également identiques à celles relevées par la figure 1. Nous pouvons cependant signaler que les voyelles ouvertes /a/, /ɛ/ et /ɔ/ mesurées en syllabe finale de mots ont des valeurs de F1 significativement plus élevées que ces mêmes voyelles mesurées en syllabe initiale de mots. Ces différences observées sur l'axe F1 ont également été notées par Gendrot [4] sur des triangles vocaliques effectués sur la base des catégories de f0 mentionnées ci-dessus et non plus de catégories de durée.

Les résultats diffèrent également concernant la position de la syllabe dans le mot pour les voyelles fermées /i/, /y/, /u/ et /o/. En effet en syllabe finale de mot, les variations formantiques pour ces voyelles sur l'axe F1 sont non significatives ou peu importantes. A l'inverse en syllabe initiale, un net détachement du 1^er formant des voyelles fermées courtes (avec des valeurs plus élevées) par rapport à leurs contreparties plus longues peut être observé. Nous suggérons ici que les variations observées dans la distribution de ces voyelles, ainsi que les variations formantiques mesurées correspondent à la répartition naturelle de ces voyelles. La position allongeante de fin de mot implique donc des voyelles plus périphériques, à l'exception des voyelles antérieures fermées /i/ et /e/. Les variations formantiques de ces voyelles avaient été notées comme plus faibles dans Gendrot & Adda-Decker [2].

4. VARIATIONS FORMANTIQUES AUTOUR DES PAUSES

Nous avons également réalisé des mesures identiques en fonction de la présence ou non de pauses immédiatement avant et après les voyelles analysées. Les pauses ont été déterminées dans la transcription par l'algorithme d'alignement suivant le même principe que les phonèmes (voir Gendrot & Adda-Decker [2]). La table 3 résume ainsi les catégories obtenues au moyen d'exemples, ainsi que le nombre d'occurrences recueillies.

Table 3 : Résumé des différentes catégories sélectionnées

	pause précédant V	pause suivant V
pause	**[pause] a**bri	matel**a**s **[pause]**
occur-rences	avec:920 (dur moy = 100ms) sans:22000 (dur moy=70ms)	avec :1600 (dur moy=120ms) sans:21300 (dur moy= 68ms)

Rappelons ici qu'il est particulièrement délicat de se baser sur la syntaxe de la transcription pour déterminer des éventuelles frontières prosodiques comme cela est fait pour l'analyse de phrases lues. En effet, il est fréquent par exemple de constater la présence de pauses entre un article et le nom qui lui est associé, ce qui va à l'encontre de bon nombre de prédictions, et est souvent considéré comme une (pause de) mise en valeur du mot suivant cette pause. Nous avons ainsi décidé de déterminer des frontières prosodiques grâce à la détection de pauses. En effet, la présence d'une pause en français est fréquemment corrélée à une fin de groupe intonatif qui est également marquée par un fort allongement vocalique.

Notre principal but était de détecter un allongement vocalique qui pouvait se produire dans l'entourage de frontières syntaxiques et/ou prosodiques. Comme cela est illustré par les figure 4a et 4b (en Annexes), les voyelles précédées ou suivies d'une pause sont caractérisées par des valeurs plus extrêmes, occupant ainsi un espace acoustique plus large, comme nous l'avions noté pour les voyelles les plus longues sur la figure 1. Ces résultats seront plus longuement développés dans la partie suivante.

5. DISCUSSION ET CONCLUSION

Il est intéressant de tenter de relier ces résultats à des travaux récents de prosodie articulatoire. En effet, une question fondamentale de cet axe de recherche est l'analyse supra-glottique des voyelles en termes des hypothèses de l'expansion de la sonorité (Straka [5]; Beckman et al. [6]) ou du renforcement des traits distinctifs (de Jong [7]). La seconde hypothèse peut être différenciée de la première sur la base de mesures de formants des voyelles fermées telles que /i/ et /u/. Selon la première hypothèse, une voyelle antérieure fermée /i/ « renforcée » sera plus ouverte (de par la mandibule et/ou les lèvres) avec des valeurs plus élevées du 1^er formant, bien que cette ouverture du conduit vocal soit en partie contradictoire avec le trait fermé du /i/. Pour la seconde hypothèse, la voyelle antérieure fermée /i/ « renforcée » sera encore plus fermée (F1 plus bas), plus antérieure et plus étirée (F2 et si plus antérieur, F3 plus élevé, une fréquence plus élevée du 3ème formant semblant être une caractéristique d'un /i/ bien formé en français).

Fougeron [8] a montré que les phonèmes en début de constituant prosodique de haut niveau tel que le groupe intonatif (GI) tendent à être hyperarticulés, ce que l'on pourrait relier ici à l'hypothèse de renforcement des traits distinctifs et que nous observons pour notre corpus (fig. 4a) à l'exception de la voyelle /i/ cependant. Nous avons pu observer cependant qu'une grande proportion (80%) de ces /i/ sont représentés par le pronom « il », qui est plus vraisemblablement réalisé avec une aperture plus grande dans cette position. Quant aux voyelles suivies d'une pause, Tabain [9] a montré que les voyelles finales (principalement /a/) de constituants prosodiques de haut niveau tels que le groupe intonatif tendaient également à être renforcées, ce que nous observons ici sur la figure 4b : lorsqu'une voyelle est suivie d'une pause, ses valeurs formantiques sont plus extrêmes sur les axes F1/F2. A nouveau, ces variations sont plus faibles pour les voyelles antérieures fermées.

De par la faible quantité de voyelles recueillies au contact immédiat d'une pause, nous ne pouvons découper ces voyelles en catégories de durée comme réalisé

précédemment. Malgré tout, des mesures effectuées sur le corpus ESTER ont permis de mesurer des variations formantiques significatives en fonction de la durée dans l'entourage d'une pause, du même ordre que celles observées pour l'ensemble des voyelles.

Le résultat principal de cette étude est la mise en évidence de la périphéralité des voyelles françaises dans deux positions prosodiques: (i) en syllabe finale de mot (comparées aux voyelles en syllabe initiale de mot (ii) dans l'entourage immédiat (avant et après) de la pause. Nous avons également montré que les voyelles en syllabe finale et initiale de mot révèlent des variations de durée induisant des variations formantiques similaires à celles observées sur l'ensemble des voyelles, malgré les différences que nous avons signalées.

BIBLIOGRAPHIE

[1] Gauvain, J.L., Lamel, L. & Adda, G. The Limsi Broadcast News Transcription System, Speech Communication, 37(1-2):89-108, 2002.

[2] Gendrot, C. & Adda-Decker, M. Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German.. Eurospeech – Lisbon (Portugal), September 2005, pp2453-2456, 2005.

[3] Delattre, P. Comparing the prosodic features in English, German, Spanish, and French. Int. Rev. Applied. Linguistics I, 193-210. volume 4, pages 2379-2382, 1965.

[4] Gendrot, C. Aspects perceptifs, physiologiques et acoustiques de différentes catégories prosodiques en français, Thèse de doctorat, 2005.Université de Paris3-Sorbonne Nouvelle.

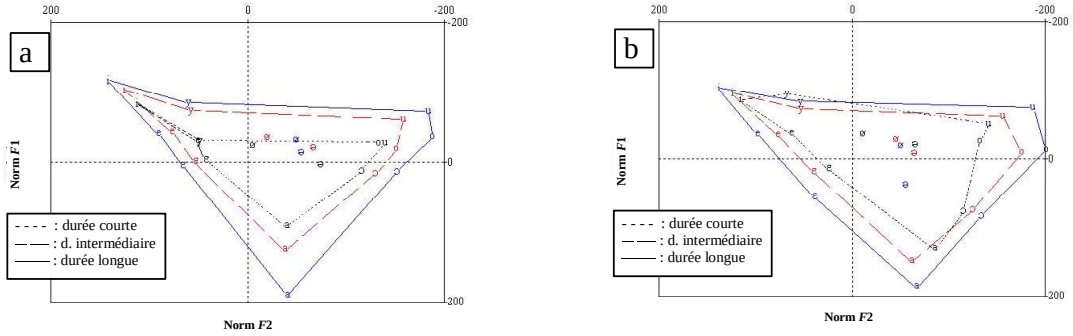
[5] Straka G. La division des sons du langage en voyelles et consonnes peut-elle être justifiée?, Travaux de Linguistique et de littérature, Université de Strasbourg, 1, pp. 17-99, 1963.

[6] Beckman M. E., Edwards J. & Fletcher J. Prosodic Structure and Tempo in a Sonority Model of Articulatory Dynamics, in Papers in lab. phon. II. Cambridge (England), Cambridge University Press, pp.68-89, 1992.

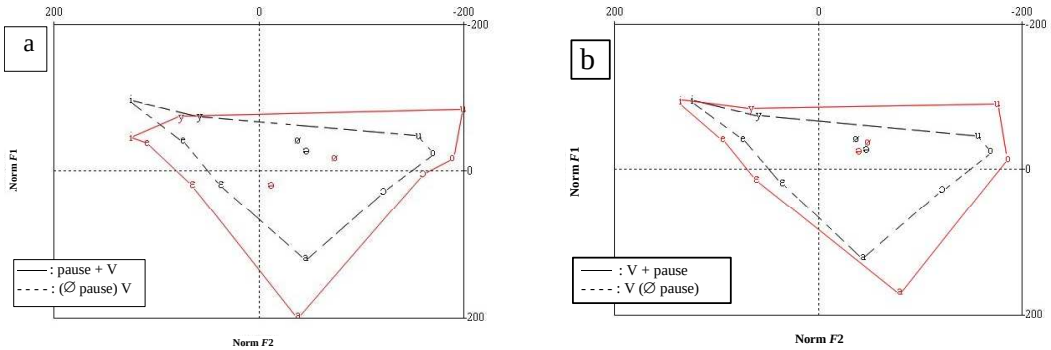
[7] de Jong K. J. The supraglottal articulation of prominence in English, *JASA*, 97, pp.491-504, 1995.

[8] Fougeron C. Articulatory properties of initial segments in several prosodic constituents in French, Journal of Phonetics, 29(2), pp. 109-135, 2001.

[9] Tabain M. Effects of prosodic boundary on /aC/ sequences: acoustic results, JASA, 113, pp. 516-531, 2003.



Figures 3ab : Valeurs moyennes de F1 et F2 pour les voyelles en fonction de leur durée (Normalisation lobanov). **a(gauche) : voyelles en syllabe initiale** de mots **b(droite) : voyelles en syllabe finale** de mots.



Figures 4ab: Valeurs moyennes de F1 et F2 pour les voyelles en fonction de l'absence/présence de la pause à proximité de la voyelle analysée. **a(gauche) : pause précédant la voyelle** ; **b(droite) : pause suivant la voyelle**.

Les systèmes vocaliques des dialectes de l'anglais britannique

Emmanuel Ferragne, François Pellegrino

Laboratoire Dynamique Du Langage
UMR CNRS 5596 – Université Lyon 2
14 avenue Berthelot
69007 LYON
Emmanuel.Ferragne@univ-lyon2.fr
http://www.ddl.ish-lyon.cnrs.fr

ABSTRACT

This paper is an attempt to characterize the vowel systems of the dialects of British English. We carried out a (semi-) automatic dialect identification procedure using [hVd] words. Our second aim was to examine to what extent the procedure allowed the description of vowel systems. The method yields approximately 90% correct identification, and we show that it is not sensitive to gender differences, and may therefore be used for the description of vowel systems.

1. INTRODUCTION

Les dialectes de l'anglais britannique ont été abondamment décrits dans la littérature traditionnelle d'avant l'ère de la phonétique de corpus (Wells [1]). Les ouvrages de référence récents, quoique basés sur l'analyse minutieuse de corpora recueillis *ad hoc* (Kortmann & Schneider [2], Foulkes & Docherty [3]), se concentrent davantage sur la variation sociolinguistique à l'intérieur d'un même dialecte, si bien que la mise en lumière des caractéristiques de prononciation régionales actuelles – au-delà de certains schibboleths bien connus – n'est pas aisée. Il semblerait même que les études contemporaines aient tendance à considérer la variation sociolinguistique seule comme digne d'intérêt, reléguant la variation régionale au statut de curiosité folklorique (voir par exemple le projet *Voices* de la BBC[1]).

Les bases de données d'enregistrements des dialectes britanniques récentes et disponibles pour la recherche sont, à notre connaissance, au nombre de deux : le corpus IViE[2], conçu pour l'étude de l'intonation, et le corpus Accents of the British Isles (ABI – D'Arcy *et al.* [4]), que nous utiliserons dans cette étude.

Notre analyse consiste à procéder à une identification (semi-) automatique du dialecte basée sur l'information contenue dans les systèmes vocaliques des variétés régionales de notre corpus. Parallèlement à l'identification automatique, nous aborderons des aspects plus phonologiques, notamment, les convergences et scissions phonémiques.

[1] http://www.bbc.co.uk/voices/
[2] http://www.phon.ox.ac.uk/~esther/ivyweb/

Les différences systémiques les plus diagnostiques en Angleterre proviennent de la scission phonémique de FOOT-STRUT et du phénomène appelé BATH-Broadening. Le premier fait référence au fait qu'avant le 17[e] siècle, *luck* et *look* étaient de parfaits homophones (avec un timbre proche de [u]). A cette époque, une délabialisation de la voyelle de *luck*, amorcée dans le sud de l'Angleterre a conduit à une phonémisation de l'opposition FOOT-STRUT au sud (Wells [1]). Ce phénomène ne s'est pas diffusé au nord, si bien que l'absence de scission y constitue un trait caractéristique, souvent stigmatisé. Le BATH-Broadening est également une innovation du sud de l'Angleterre ; la voyelle s'est allongée devant fricative au 17[e] siècle et a acquis une qualité postérieure au 19[e]. Certes, le statut phonémique de l'opposition BATH-TRAP demeure discutable – au reste, cette opposition n'a pas pu être étudiée dans notre corpus – mais elle méritait d'être mentionnée car elle constitue, au même titre que FOOT-STRUT une frontière nord/sud bien connue du dialectologue et du locuteur moyen. Wells [1] et Beal [5] notent l'existence d'une convergence NURSE-SQUARE à Liverpool ; c'est également le cas à Hull (Williams & Kerswill [6]), ville où ont eu lieu les enregistrements du dialecte étiqueté "East Yorkshire" dans notre corpus. Une description sommaire de l'anglais écossais fait apparaître une absence de contraste possible entre les voyelles de GOOSE et FOOT et celles de LOT et THOUGHT. Un examen plus détaillé et récent (Stuart-Smith [7]) tend à confirmer ces convergences phonémiques (au moins à Glasgow) tout en soulignant qu'elles varient en fonction de facteurs sociolinguistiques. FOOT et STRUT sont, quant à eux bien distincts, et le BATH-Broadening n'a pas eu lieu. L'anglais du Pays de Galles, tel qu'il est décrit chez Wells [1] et Penhallurick [8] ne fait pas apparaître de convergence ou scission phonémique que nous puissions tester. En effet, Wells [1] mentionne la convergence STRUT-Schwa (ici encore, le statut phonémique est contestable), ce que nous ne pouvons pas étudier puisque nous ne disposons que de voyelles accentuées et que les membres de l'ensemble lexical Schwa apparaissent en position inaccentuée. Penhallurick [8] mentionne toutefois la possibilité d'une réalisation [u] pour STRUT dans le nord-est du pays, ce qui nous laisserait peut-être envisager l'existence d'une convergence FOOT-STRUT.

2. CORPUS ET MÉTHODE

2.1. Description du corpus

Nous avons utilisé les mots en [hVd] du corpus Accents of the British Isles (ABI). Le corpus ABI est une base d'enregistrements de parole lue effectués en 2003 et qui regroupe 14 dialectes des Iles Britanniques, pour un total de 284 locuteurs. En moyenne, chaque dialecte est représenté par dix locuteurs masculins et dix locutrices féminines. Chaque locuteur répète cinq fois une liste de 19 mots, dont certains sont des non-mots, à structure [hVd]. Chacune des 19 voyelles employées constitue une opposition potentielle dans le système phonologique. En d'autres termes, les dialectes possédant l'inventaire vocalique le plus riche ont jusqu'à 19 phonèmes. Les fichiers son sont au format Windows PCM, 22050 Hz, 16 bits, mono. Les enregistrements ont eu lieu dans des salles calmes (ex : salles de bibliothèques publiques) avec un micro-casque directement relié à un PC via une carte son externe. La liste des dialectes et les abréviations correspondantes est donnée dans la table 1. Ces enregistrements sont majoritairement de qualité moyenne en cela qu'ils comportent certains bruits environnants, des interventions de la personne qui effectue l'enregistrement, et que la compétence des locuteurs de certains dialectes en lecture est – aux dires des phonéticiens anglophones qui ont eu l'occasion de les entendre – très décevante. De plus, si l'on se place dans une optique de dialectologie moderne (dominée par la sociolinguistique), le corpus est inexploitable puisque aucune donnée n'est fournie concernant l'âge des locuteurs, leur occupation et leur histoire linguistique.

Table 1 : liste des dialectes.

abréviation	étiquette du dialecte
brm	Birmingham
crn	Cornouailles
ean	East-Anglia
eyk	East-Yorkshire
gla	Glasgow
lan	Lancachire
lvp	Liverpool
ncl	Newcaslte
nwa	North Wales
roi	Republic of Ireland
shl	Scottish Highlands
sse	Standard Southern English
uls	Ulster

2.2. Pré-traitement des données

Le corpus ABI est segmenté au niveau du mot. A partir des frontières délimitant chaque mot en [hVd], nous avons procédé à une détection automatique du voisement (avec la librairie Snack en Tcl/Tk[3]) ; nous avons alors considéré que ces trames voisées constituaient la partie vocalique du mot. A partir des voyelles ainsi isolées, 12 MFCC plus l'énergie ont été calculés à l'aide de la toolbox Rasta[4] pour Matlab avec un pas d'analyse de 10 ms, une fenêtre de 20 ms. Chaque voyelle est ensuite représentée par un vecteur de 52 valeurs : 12 MFCC et l'énergie à ¼, ½ et ¾ de la durée, les deltas MFCC et la dérivée de l'énergie calculés sur 5 trames centrées sur la trame du milieu temporel ainsi que, optionnellement, une 53[e] valeur : la durée de la voyelle.

2.3. Classification

La méthode de classification utilisée ici s'inspire largement des travaux de Huckvale [9]. Elle consiste à obtenir une représentation du système vocalique de chaque locuteur à partir d'une matrice de distances euclidiennes entre les voyelles de ce locuteur. Ceci constitue en soi une forme de normalisation du locuteur puisque l'on ne compare pas directement les voyelles d'un locuteur avec celles d'un autre, mais les voyelles d'un même locuteur entre elles. Pour un même locuteur et un timbre donné, ce sont les valeurs moyennes qui ont été employées pour l'analyse et non pas les valeurs individuelles de chaque répétition. On calcule donc la matrice de distances entre N voyelles pour chaque locuteur, puis la matrice de distance moyenne pour chaque dialecte. La méthode de validation est basée sur le principe du "leave-one-out" : la matrice de distances entre segments de chaque locuteur est comparée à la matrice (moyenne) de chaque dialecte – dont celle de son dialecte calculée sans la matrice du locuteur en cours de test. Ce dernier est classé comme appartenant au dialecte dont la matrice se rapproche le plus de la sienne. La similarité entre matrices est estimée grâce au coefficient de corrélation de Mantel. L'utilisation d'une corrélation entre matrices rend cette mesure insensible aux distances absolues entre les voyelles d'un locuteur. En outre, l'un des avantages de cette méthode de classification réside dans le fait qu'elle nécessite un temps de calcul extrêmement bref (moins de 7 secondes en temps CPU pour 145 locuteurs hommes sur un PC équipé de Windows XP et d'un processeur Intel Pentium 4, 3.2 GHz).

2.4. Pouvoir discriminant des différents timbres

Nous avons également souhaité tester le potentiel d'un nombre plus restreint de voyelles afin de faire apparaître les timbres les plus discriminants. Pour ce faire, la méthode de classification a été appliquée à chaque combinaison de 3 voyelles parmi 19, puis, la combinaison donnant le taux d'identification du dialecte le plus élevé a été conservée, et enrichie d'une des 16 voyelles restantes ; après avoir testé ces 16 ensembles de 4 voyelles, on garde le meilleur et on forme à partir de ce dernier 15 nouveaux ensembles, et ainsi de suite jusqu'à qu'il ne subsiste qu'un ensemble de 19 voyelles.

[3] http://www.speech.kth.se/snack/

[4] http://www.ee.columbia.edu/labrosa/matlab/rastamat/

Notons enfin que nous avons fait juger la typicalité de chacun des 145 locuteurs hommes du corpus sur la base d'un passage lu d'environ 90 mots par un phonéticien britannique expert. Cette analyse auditive préalable nous a conduit à exclure le dialecte portant l'étiquette "Inner London" car il était constitué de locuteurs immigrés d'ascendances très variées formant un tout trop hétérogène pour être considéré comme une seule et même entité.

3. RÉSULTATS ET DISCUSSION

3.1. Le système vocalique à onze monophtongues

En première approximation, nous avons classé les dialectes (sans tenir compte de la durée des voyelles) à partir des onze monophtongues de l'anglais britannique standard. Les trois modèles – hommes, femmes, sexes confondus – donnent respectivement des taux d'identification du dialecte de 80,7 ; 80,5 et 84%. Nous savons que la durée vocalique joue un rôle très important dans le système de la plupart des dialectes de l'anglais britannique (sauf, très vraisemblablement en Ecosse, voir Wells [1]) ; nous avons donc ajouté la durée aux paramètres spectraux précédents et avons pondéré empiriquement la durée par rapport à ces derniers. L'inclusion de la durée dans les modèles permet d'atteindre, 83,7 ; 87,5 et 87,8% dans les conditions homme, femmes, et sexes confondus respectivement.

3.2. Prise en compte des diphtongues

Nous avons ensuite tenté de déterminer les meilleures combinaisons de p = 3 à 19 voyelles parmi 19 (voir 2.4 pour les méthodes ; la durée n'est pas incluse). Les taux d'identification en fonction du nombre de voyelles sont donnés dans la figure 1.

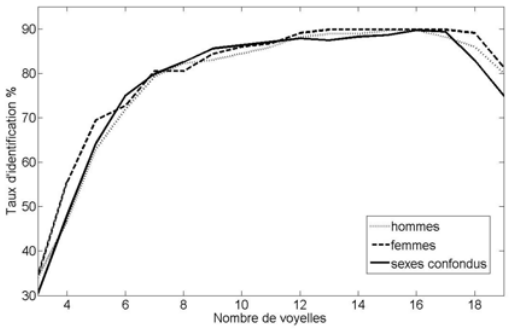


Figure 1 : taux d'identification du dialecte en fonction du nombre de voyelles incluses dans le modèle.

Le détail des meilleures combinaisons de voyelles apparaît dans la table 2. Les meilleurs taux pour les hommes, les femmes et sexes confondus sont de 89,6 ; 89,8 et 89,7% avec, respectivement, 15, 15 et 16 voyelles.

Table 2 : liste incrémentale des meilleures combinaisons de p = 3 à 19 voyelles parmi 19.

nb de voyelles	3	4	5	6	7	8	9	10	11	12	13	14	15	16
hommes	hid, hood, Hudd	heard	hide	whod	hade	hoed	hard	heed	hoard	hared	hod	head	had	*
femmes	hid, heard, hide	had	Hudd	hod	howd	head	whod	heered	hoid	hured	heed	hoard	hard	*
sexes confondus	hid, hood, Hudd	heard	hide	hade	whod	hoed	howd	hured	heered	heed	hod	head	hoard	had

Nous avons enfin inclus la durée et l'avons pondérée pour tenter d'améliorer les taux d'identification donnés dans la figure 1 à partir des combinaisons de voyelles de la table 2. Les taux d'identification ainsi obtenus sont de 89,6 ; 91,4 et 91,6% pour les conditions hommes, femmes et sexes confondus respectivement. La prise en compte de la durée ne permet d'améliorer le taux d'identification des femmes et des deux sexes confondus que dans une moindre mesure ; elle n'a en revanche aucun effet sur le modèle des hommes : cela est probablement dû à l'utilisation de mots en [hVd] hyper-articulés.

3.3. Discussion

Nous citerons, à titre de comparaison, deux études qui présentent des valeurs de formants pour l'identification du dialecte. Huckvale [9], qui a travaillé sur une autre partie du corpus que nous utilisons ici, partage ses voyelles en deux, utilise la valeur médiane des quatre premiers formants dans chaque moitié, et ses valeurs sont ensuite centrées-réduites par locuteur. Le vecteur moyen d'un timbre donné pour chaque locuteur et comparé aux vecteurs moyens de ce même timbre pour chaque dialecte. Dans la condition "tous sexes", il obtient 71,9% d'identification. Barry *et al.* [10] utilisent une méthode de classification basée sur des critères phonétiques et obtiennent environ 74% de classification correcte pour leurs 58 locuteurs en 4 aires dialectales grossières.

Dans la figure 1, nous constatons que la courbe d'identification tend à se stabiliser lorsque 7 voyelles sont prises en compte dans le modèle. Parmi ces 7 voyelles, la table 2 révèle que la paire *hood-Hudd*, qui correspond à la scission FOOT-STRUT mentionnée plus haut comme trait extrêmement diagnostique, est en bonne position pour les hommes et tous sexes confondus. La présence de *hade* n'est pas fortuite non plus puisque la diphtongaison (Ferragne & Pellegrino [11]) et un trait hautement discriminant. Le fait de retrouver cette voyelle aussi bien placée dans la table 2 tend à prouver que notre méthode capture de façon adéquate les caractéristiques dynamiques de voyelles.

Notons encore qu'un des traits phonétiques que nous avons mesuré n'est pas à proprement parler vocalique : il s'agit de la rhoticité, i.e, la réalisation ou non d'un /r/ postvocalique, et la nature articulatoire de ce /r/. Par exemple, le sud-est de l'Angleterre n'est pas rhotique, le sud-ouest a un /r/ souvent rétroflexe (Altendorf & Watt [12] ; au Pays de Galles, le /r/ peut être une approximante post-alvéolaire, un trille ou une battue (Penhallurick [8]). Il est donc certain que, outre les différences des systèmes vocaliques, notre

classification a bénéficié de la variation géographique du /r/.

Comme l'a noté Huckvale [9], nos résultats confirment que cette méthode est insensible aux différences de sexe. Elle semble d'ailleurs tout a fait adéquate pour représenter des systèmes vocaliques en évitant d'avoir recours à une normalisation du locuteur fastidieuse et souvent inefficace basée sur des valeurs de formants. L'utilisation des MFCC permet une procédure entièrement automatisée alors que l'extraction automatique de formants nécessite souvent l'intervention d'un expert humain. Afin d'illustrer l'intérêt de notre méthode pour la description phonétique, la figure 2 représente un dendrogramme obtenu à partir de la matrice de distance moyenne des 11 monophtongues du dialecte East Yorkshire. On visualise aisément l'absence de scission FOOT-STRUT (*hood* et *Hudd* sur la figure), typique des dialectes du nord de l'Angleterre, et la proximité relative des voyelles de *heed* et *hid*, de *head* et *heard*, de *had* et *hard*, et de *hod* et *hoard*.

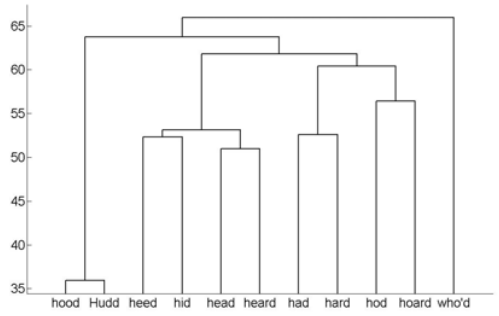


Figure 2 : Classification hiérarchique des voyelles du dialecte East Yorkshire.

En ce qui concerne les résultats de l'identification, nous pouvons les comparer directement avec ceux de Huckvale [9] : ce dernier avait obtenu 86,9% dans la condition tous sexes confondus, et 87,2% pour la condition "même sexe" en comparant des matrices de distances de 140 voyelles issues de phrases lues. Il est intéressant de noter que malgré le fait que les mots en [hVd] soient communément considérés comme trop éloignés de la parole spontanée, notre étude prouve qu'ils contiennent suffisamment d'information pour la description et l'identification de certains dialectes.

4. CONCLUSION

A partir des MFCC extraits des mots en [hVd] du corpus ABI, nous obtenons un taux d'identification du dialecte de 89,6 ; 91,4 et 91,6% pour les hommes, femmes et tous sexes confondus respectivement. Le taux d'identification à partir des MFCC des monophtongues (84%) est amélioré lorsque la durée est incluse dans le modèle (87,8%), et l'ajout des diphtongues permet d'atteindre 91,6%. La méthode des matrices de distances entre timbres est ainsi robuste par rapport aux différences de sexe. Ceci permet une représentation fiable des systèmes vocaliques des différents dialectes de l'anglais britannique.

BIBLIOGRAPHIE

[1] J.C. Wells. *Accents of English*. Cambridge University Press, Cambridge, UK, 1982.

[2] B. Kortmann and E.W. Schneider. *A Handbook of Varieties of English*. Mouton de Gruyter, Berlin, 2004.

[3] P. Foulkes and G. Docherty. *Urban Voices. Accent Studies in the British Isles*. Arnold, Londres, 1999.

[4] S.M. D'Arcy, M.J. Russell, S.R. Browning and M.J. Tomlinson. The Accents of the British Isles (ABI) corpus. *In Proc. Colloque MIDL 2004*, pages 115-119, 2004.

[5] J. Beal. English dialects in the North of England: phonology. *In A Handbook of Varieties of English*, B. Kortmann and E.W. Schneider (eds), pages 113-133, Cambridge University Press, UK, 2004.

[6] A. Williams and P. Kerswill. Dialect levelling: change and continuity in Milton Keynes, Reading and Hull. *In Urban Voices. Accent Studies in the British Isles*, P. Foulkes and G. Docherty (eds), pages 141-162, Arnold, Londres, 1999.

[7] J. Stuart-Smith. Scottish English: phonology. *In A Handbook of Varieties of English*, B. Kortmann and E.W. Schneider (eds), pages 47-67, Cambridge University Press, UK, 2004.

[8] R. Penhallurick. Welsh English: phonology. *In A Handbook of Varieties of English*, B. Kortmann and E.W. Schneider (eds), pages 98-112, Cambridge University Press, UK, 2004.

[9] M. Huckvale. ACCDIST: a Metric for Comparing Speakers' Accents. Proc. Interspeech 2004 ICSLP, 2004.

[10] W.J. Barry, C.E. Hoequist and F.J. Nolan. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, 3: 355-366, 1989.

[11] E. Ferragne and F. Pellegrino. Diphthongization as a cue for the automatic identification of British English dialects. *In Proc. 148th Meeting of the Acoustical Society of America*, 2004.

[12] U. Altendorf and D. Watt. The dialects in the South of England: phonology. *In A Handbook of Varieties of English*, B. Kortmann and E.W. Schneider (eds), pages 178-203, Cambridge University Press, UK, 2004.

Session XVI

Poster

Jeudi 15 juin 2006 - 11h15 12h30

Reconnaissance audiovisuelle de la parole par VMike

Fabian Brugger [1], *Leila Zouari* [1], *Hervé Bredin* [1], *Asmaa Amehraye* [2], *Gérard Chollet* [1], *Dominique Pastor* [2] *et Yang Ni* [3]

[1] GET-ENST/CNRS-LTCI - LastName@tsi.enst.fr - 46 rue Barrault 75634 Paris cedex 13, France
[2] GET-ENST Bretagne - FirstName.LastName@enst-bretagne.fr - Technopôle Brest-Iroise - 29238 Brest Cedex 3
[3] GET-INT - FirstName.LastName@int-evry.fr - 9 rue Charles Fourier 91011 Évry cedex

ABSTRACT

This article presents a new Electronic Retina based Smart Microphone (VMike) and investigates the use of its novel parameters - lip profiles - in audiovisual speech recognition. In order to evaluate the parameterization, both an audio only and a video only speech recognition system are developed and tested. Then, two main fusion techniques are employed to test the usability of profiles in audiovisual systems : feature fusion and decision fusion. These results are compared to the performance of recognizers based on a state-of-the-art parameterization, and also to results obtained by applying perceptual filtering to the speech signal prior to recognition. When feature fusion is applied, and under noisy conditions, recognition using lip profiles improved by up to 13 percent with respect to audio-only recognition.

1. INTRODUCTION

Les systèmes actuels de reconnaissance de la parole sur petit vocabulaire et dans un environnement non bruité donnent des résultats satisfaisants (taux d'erreur inférieur à 1 %) [8]. Cependant, dans les applications réelles, les conditions d'enregistrement (voiture, avion, hélicoptère, etc.), d'acquisition et de transmission ne sont pas idéales : un bruit est nécessairement introduit dans la chaîne de traitement de la parole, entraînant une baisse des performances du système de reconnaissance.
Plusieurs méthodes ont été développées dans le but d'améliorer la reconnaissance de la parole en milieu bruité. On distingue trois principales catégories : le débruitage du signal, l'adaptation du système à l'environnement ou encore la reconnaissance audiovisuelle en considérant le mouvement des lèvres. Cette dernière méthode repose sur l'idée que la parole est un moyen audiovisuel de communication. En effet, le message parlé est plus intelligible quand il est accompagné de la vision du visage du locuteur, et particulièrement quand le milieu de transmission est bruité.
Un système de reconnaissance audiovisuelle de la parole résulte de la fusion de deux systèmes mono-modaux audio et vidéo. Classiquement, on distingue deux types de fusion : la fusion des paramètres et la fusion des scores.
Dans le cadre du projet VMike, un dispositif éponyme a été développé. Il s'agit d'un microphone augmenté d'une rétine électronique produisant un signal de parole audiovisuelle. Le but de cet article est de montrer que l'utilisation de cette nouvelle paramétrisation de la zone de la bouche -issue de VMike- peut être utilisée pour améliorer les taux de reconnaissance de la parole en milieu bruité.

Ce document est organisé comme suit. La section 2 présente rapidement le dispositif VMike ainsi que les paramètres visuels qu'il produit. Très peu de données ont été acquises avec le dispositif lui-même : nous les avons donc simulé à partir de la base de données audiovisuelles BANCA. Les paramètres utilisés sont décrits dans la section 3. Un tour d'horizon des techniques de fusion expérimentées constitue la section 4. Les expériences réalisées sont présentées en détails dans la section 5, où nous comparons notre système à un système audiovisuel état-de-l'art basé sur une transformation DCT de la zone des lèvres et à un système *audio seul* basé sur un filtrage perceptuel. Enfin, la section 6 conclut cet article et propose différentes perspectives et pistes d'amélioration.

2. VMIKE

2.1. Description du dispositif

VMike est un microphone augmenté d'une rétine électronique. Il a été développé dans le cadre du projet VMike[1] afin d'explorer l'apport des lèvres en reconnaissance de la parole. Les spécifications de ce dispositif (schématisé dans la figure 1) sont :

– un canal *voix* assuré par un microphone classique,
– un canal *vision* assuré par une rétine électronique.

Cette rétine permet la compression de l'information visuelle par projection sur les axes horizontal et vertical. L'interface stéréo entraîne une synchronisation parfaite et naturelle entre les deux signaux audio et vidéo. Les avantages d'une telle configuration sont :

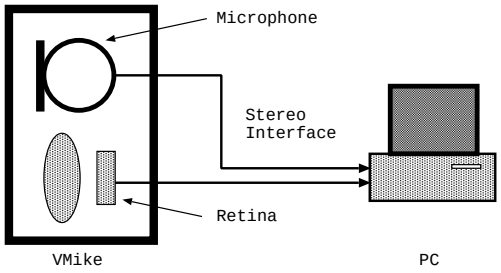


FIG. 1: le dispositif VMike

[1]Cette étude est soutenue par le groupe des Ecoles de Télécommunications (GET)

– sa simplicité,
– le nombre réduit de paramètres qu'elle produit,
– son importante vitesse de transmission.

2.2. *Sorties du VMike*

Deux signaux mono sont générés par le dispositif VMike :
– le signal audio de parole sur le canal gauche,
– les projections horizontales et verticales de la zone des lèvres sur le canal droit.

Alors que la parole est acquise de façon classique par le microphone, les images subissent un prétraitement avant leur transmission. En effet, pour chaque image, les projections sur les axes horizontal et vertical (voir figure 2) sont calculées (et concaténés), modulées puis transmises sur le canal droit.

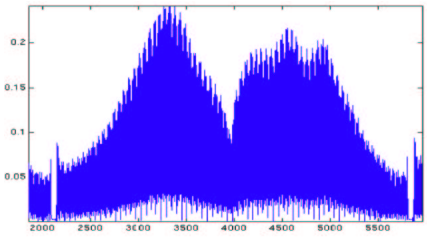


FIG. 2: Exemple d'une sortie vidéo démodulée issue du dispositif VMike

3. PARAMÈTRES VISUELS

3.1. *Simulation du VMike*

Au moment de la réalisation des expériences dont les résultats sont présentés dans cet article, peu de données ont été enregistrées avec le dispositif VMike. Aussi, nous avons simulé des enregistrements à partir de la base de données audiovisuelles BANCA [3].
Rappelons que le VMike est utilisé comme un microphone classique, c'est-à-dire tenu à quelques centimètres de la bouche et orienté vers celle-ci. Aussi avons-nous implémenté un algorithme (présenté en détails dans [4]) permettant la localisation automatique de la zone de la bouche dans les séquences vidéo de la base de données BANCA. Un exemple de cette zone est présenté dans la figure 5.

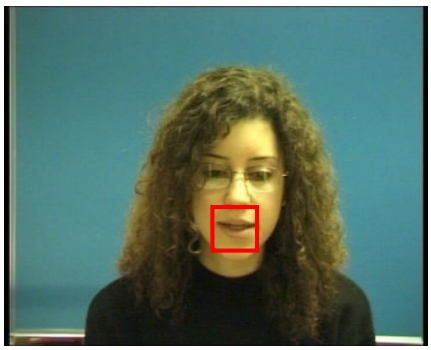


FIG. 3: Détection de la zone d'intérêt

3.2. *Projections XY*

Une fois la zone de la bouche localisée, elle est normalisée en taille (200 par 200) et les projections sur l'axe horizontal (respectivement vertical) sont calculées très simplement comme la somme des niveaux de gris sur chaque colonne (respectivement chaque ligne). La figure 4 illustre cette transformation.

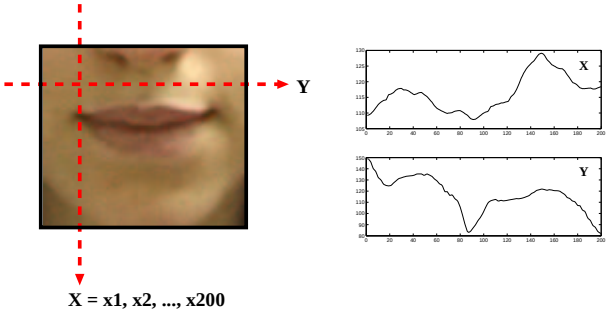


FIG. 4: Les projections de Vmike

3.3. *DCT*

Afin de comparer l'apport de cette nouvelle approche, une paramétrisation état-de-l'art [9] a aussi été implémentée. La zone de la bouche est normalisée en taille ($H = 64$ pixels par $W = 64$ pixels). Notant $I(i,j)$ l'intensité des pixels dans cette zone, une DCT (Discret Cosine Transform) est appliquée sur la zone d'intérêt :

$$X(u,v) = C(u,v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \alpha(i,u,H)\, \alpha(j,v,W)\, I(i,j)$$

où $C(u,v)$ est un coefficient de normalisation et $\alpha(i,u,H) = \cos\left(\frac{2i+1}{2H} u\pi\right)$. Afin de réduire la dimension de ces paramètres (initialement $64 \times 64 = 4096$), les indices des coefficients les plus énergétiques sont obtenus sur un ensemble d'apprentissage. Seuls les coefficients correspondants sont conservés : dans notre cas, nous avons conservé les 100 coefficients les plus énergétiques. Ce principe est utilisé classiquement dans le cadre de la compression d'images : ils correspondent généralement aux fréquences u et v les plus basses.

3.4. *Traitement des paramètres bruts*

Plusieurs transformations sont alors appliquées à ces paramètres dans le but de réduire leur dimension, augmenter leur pouvoir discriminant et tenir compte de leur caractère dynamique.
Ainsi, une normalisation de la moyenne sur les paramètres bruts est suivie d'une transformation en analyse discriminante linéaire (LDA) pour réduire la dimension à 40 : ce sont les paramètres DCT et Pro (projections) dans la suite de cet article.
La dynamique de ces paramètres est modélisée en concaténant, à chaque instant d'échantillonnage, 15 échantillons consécutifs centrés sur l'échantillon courant. Une LDA est alors appliquée sur ces paramètres de dimension $40 \times 15 = 600$ pour obtenir des paramètres discriminants contenant l'information dynamique de dimension 40 : ce sont les paramètres DCT2 et Pro2.

4. TECHNIQUES DE FUSION

L'objectif d'un système de reconnaissance audiovisuelle est de combiner au mieux les performances de deux systèmes audio et vidéo afin d'améliorer les performances de reconnaissance de la parole, en particulier en présence de bruit. Classiquement, on distingue deux types de fusion : la fusion des paramètres et la fusion des scores.

4.1. Fusion des paramètres

Cette fusion est réalisée au moment de la paramétrisation des signaux audio et vidéo. Une fois les paramètres de chaque modalité sont extraits, les vecteurs audio $O_{a,t}$ et vidéo $O_{v,t}$, de dimension d_a et d_v respectivement, sont concaténés à chaque instant t pour ne former qu'un seul vecteur de paramètres audiovisuels $O_{av,t} = [O_{a,t}, O_{v,t}]$ de dimension $d_a + d_v$.
Dans les étapes suivantes de la chaîne de reconnaissance de la parole (estimation des paramètres, décodage, évaluation), aucune modification n'est nécessaire.

4.2. Fusion des scores

La fusion de scores ou de décision est possible lorsque l'on dispose de systèmes séparés (ici, audio et vidéo) et que leur fusion est réalisée au moment de la décision, par combinaison de leurs scores respectifs. Des poids différents peuvent être affectés à chaque système (ou parties de ces derniers) afin de privilégier l'une ou l'autre des deux modalités. Dans le cas de système de reconnaissance où les unités sub-lexicales (de type *phone*, par exemple) sont modélisées par des modèles de Markov cachés, cette fusion peut avoir lieu à différents niveaux qui sont l'état ou le phone ou le mot ou encore la phrase. Lorsque la fusion est éffectuée à chaque état, elle est dite synchrone, sinon elle est asynchrone.
Une fusion audiovisuelle synchrone est réalisée comme suit. Soient deux systèmes de reconnaissance de la parole audio et vidéo dont les modèles acoustiques ont la même topologie. Si $P(O_a; t, s)$ et $P(O_v; t, s)$ représentent les vraisemblances respectives d'une observation O émise à l'instant t par le même état s audio et vidéo respectivement, alors son score audiovisuel peut s'exprimer par :

$$P(O_{av}; t, s) = \lambda P(O_a; t, s) \times (1 - \lambda) P(O_v; t, s)$$

Le poids λ permet de donner plus d'importance à une modalité ou à l'autre. Pour chaque système, λ peut être choisi constant ou variable. Généralement, il dépend du rapport signal à bruit. Des travaux antérieurs [5] montrent que les performances du système de reconnaissance audiovisuelle sont meilleures pour un paramètre λ dynamique. .

5. EXPÉRIENCES ET RÉSULTATS

Deux types d'expérience ont été réalisées : reconnaissance audiovisuelle de la parole et débruitage par filtrage perceptuel en amont de la reconnaissance de la parole. Dans le premier cas, les paramètres vidéo de type DCT et projections sont testés. Le même protocole expérimental est adopté dans les deux cas.

5.1. Protocole expérimental

Les expériences de reconnaissance de chiffres décrites ci-dessous ont été réalisées sur les données en condition *studio* de la base de données audiovisuelle BANCA (partie *controlled*). 52 locuteurs (26 femmes et 26 hommes) ont enregistré 4 sessions (S1 à S4) de 2 phrases chacune durant lesquelles il/elle prononce une suite aléatoire de 12 chiffres (parmi les chiffres 1 à 9) suivie d'autres informations (nom, adresse, ..). Seulement les chiffres sont pris en considération dans ces expériences.
Les 6 phrases des trois premières sessions sont utilisées lors de l'apprentissage et les 2 phrases de la quatrième session pour effectuer les tests. Par conséquent, $52 \times 6 = 312$ phrases de 12 chiffres (environ 38 minutes) consituent l'ensemble d'apprentissage et $52 \times 2 = 104$ tests ont été réalisés.
Des modèles de Markov cachés (3 états chacun et 16 gaussiennes par état) indépendants du contexte sont construits et estimés par l'algorithme de Baum-Welch. Le signal de parole est paramétrisé par des vecteurs de 12 coefficients MFCC et de leurs première et deuxième dérivées. Pour la vidéo, selon les expériences, les coefficients DCT ou les projections sont utilisés. Le bruit testé est de type *babble*. Il est extrait de la base de données NoiseX [1]. Enfin, le décodage est réalisé par le décodeur HTK [10], en utilisant un algorithme de Viterbi.

5.2. Reconnaissance audiovisuelle

Comme mentionné auparavant, l'objectif de ce travail est d'expérimenter l'apport des projections des lèvres en reconnaissance de la parole bruitée. Pour ce faire, la base de données BANCA est bruitée par un bruit de type *babble*. Puis deux systèmes audio et vidéo sont construits et leurs résultats sont évalués. Suivant le type de paramètres, on obtient une précision de 37.02% pour le système vidéo utilisant les projections et 46.69% avec les DCT. Enfin, une technique de fusion (des paramètres ou des scores) permet de générer le système audiovisuel. Les résultats de reconnaissance audiovisuelle sans bruit et à -5dB sont reportés sur le tableau 1 :

TAB. 1: Résultats des systèmes de reconnaissance audio, vidéo et et audiovisuels

	Monomodal		Fusion param.		Fusion scores	
	studio	-5dB	studio	-5dB	studio	-5dB
Audio	97.55	45.59	-	-	-	-
Pro	35.49	35.49	96.71	49.93	97.55	46.01
Pro2	37.02	37.02	95.31	51.68	97.55	46.71
DCT	46.69	46.69	93.98	52.52	97.55	50.91
DCT2	44.40	44.40	95.45	54.22	97.55	51.15

On peut remarquer que :

- Dans les deux cas de fusion, la reconnaissance audiovisuelle apporte une amélioration par rapport au système audio, et ce, pour tous les types de paramétrisation (Pro, Pro2, DCT, et DCT2).
- Les résulats de la fusion des paramètres sont légèrement supérieurs à ceux de la fusion synchrone des scores.
- Les meilleurs résultats de fusion des paramètres sont obtenus avec les paramètres dynamiques (Pro2 et DCT2) qui correspondent à une augmentation relative de la précision de 13% et de 19%.

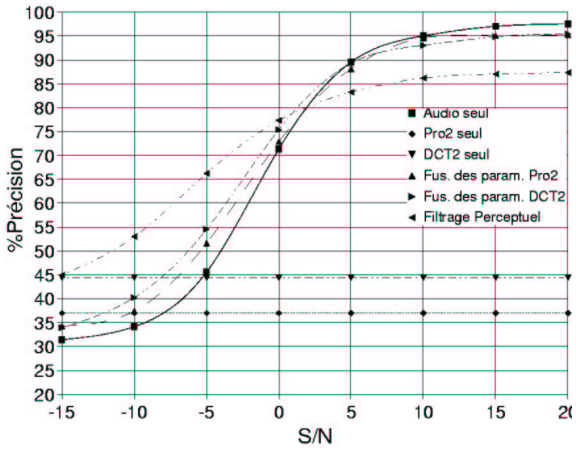


FIG. 5: Résultats de la fusion des paramètres

5.3. Débruitage de la parole

Nous avons étudié un filtrage perceptuel [6] [2] conçu de manière à respecter le phénomène de masquage simultané. Une modélisation de ce dernier permet de calculer pour chaque trame du signal de parole une courbe de masquage représentant les points de pression acoustiques nécessaires pour qu'un son devienne audible en présence d'un masquant [7]. Le but est de masquer les composantes audibles du bruit résiduel et de diminuer les distorsions du signal. Nous avons alors comparé les taux de reconnaissance de la méthode audiovisuelle à ceux obtenus par le filtrage perceptuel [6] dans les conditions idéales où la courbe de masquage est calculée à partir de la version non bruitée du signal de parole et la densité spectrale de puissance du bruit est estimée à partir d'une référence de bruit seul. Les résultats sont reportés sur la figure 5.

On peut voir que le filtre perceptuel améliore les performances du système bruité. Ces performances dépassent celle du système de reconnaissance audiovisuelle pour des rapports signal sur bruit inférieurs à 0 dB.

6. CONCLUSION ET PERSPECTIVES

Une nouvelle rétine électronique augmentée d'un microphone a été développée. Elle permet d'acquérir un signal audiovisuel de parole où audio et vidéo sont naturellement synchronisés. Le signal vidéo est constitué des projections horizontales et verticales de l'image de la zone des lèvres de l'utilisateur. L'apport de cette paramétrisation originale a été expériementé dans le cadre d'un système de reconnaissance audiovisuelle de la parole en milieu bruité sur la base de données audiovisuelles BANCA. Deux techniques de fusion audiovisuelle ont été envisagées (au niveau des paramètres et des scores) et comparées à un système audiovisuel état-de-l'art et un système audio avec débruitage par filtre perceptuel.
Bien que moins performantes que l'état-de-l'art basé sur la transformation DCT de la zone des lèvres, les projections apportent une amélioration des performances en milieu bruité comparé à un système audio seul (+13% de précision relative). La fusion niveau des paramètres donne de meilleurs résultats que celle au niveau des scores. Cependant, cette dernière pourrait être améliorée en appliquant une fusion asynchrone.
Aussi, prévoyons-nous d'appliquer des transformations de type PCA ou LDA dans l'espace des paramètres audiovisuels. En outre, il est prévu de combiner les deux approches (débruitage et fusion audiovisuelle) afin d'obtenir un système profitant de ces deux sources d'amélioration. Enfin, à plus long terme, il serait intéressant de tester l'utilisation du VMike en situation réelle (les expériences ayant ici été menées sur des données simulées).

RÉFÉRENCES

[1] The NoiseX database. http ://spib.rice.edu/spib.

[2] A. Amehraye, D. Pastor, and S. Ben Jebara. On the application of recent results in statistical decision and estimation theory to perceptual filtering of noisy speech signals. In *ISCCSP 06*, 2006.

[3] Enrique Bailly-Baillière, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belen Ruiz, and Jean-Philippe Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.

[4] Hervé Bredin, Guido Aversano, Chafic Mokbel, and Gérard Chollet. The Biosecure Talking-Face Reference System. Accepté à MMUA 2006, May 2006.

[5] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognitio. In *IEEE International Conf. ASSP (ICASSP)*, Salt Lake City-USA, 2001.

[6] Y. Hu and P. Loizou. Incorporating a psychoacoustic model in frequency domain speech enhancement. In *IEEE Signal Processing Letters*, volume 6, pages 270–273, 2004.

[7] J. D Johnston. Transform coding of audio signals using perceptual noise criteria. In *IEEE Jour. Selected Areas Commun*, volume 6, pages 9956–9963, 1998.

[8] John Makhoul and Richard Schwartz. State of the art in continuous speech recognition. In *Natl. Acad. Sci.*, pages 9956–9963, 1995.

[9] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior. Recent Advances in the Automatic Recognition of Audiovisual Speech. In *IEEE*, volume 91, pages 1306 – 1326, September 2003.

[10] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.2)*. Cambridge University Engineering Department, December 2002.

Probabilité *a posteriori*: amélioration d'une mesure de confiance en reconnaissance de la parole

Julie Mauclair, Yannick Estève, Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Le Mans, France
{mauclair,esteve,deleglise}@lium.univ-lemans.fr

ABSTRACT

This paper adresses the word posterior probability used as a confidence measure for speech recognition system. We present a new confidence measure based on the behavior of the language model back-off used during the recognition process. Merging this new confidence measure with word posterior probability allows to obtain another confidence measure which outperfoms the word posterior probability. Our experiments have been carried out on the ESTER corpora, the french evaluation campaign on french broadcast news. Using the normalized cross entropy (NCE) as an evaluation metric, experimental results on test data show a very significant improvement : whereas the word posterior probability reaches a value of 0.187, the fusion measure obtains 0.270.

1 Introduction

Les mesures de confiance servent à estimer la fiabilité d'une hypothèse de reconnaissance et donc, la fiabilité du système de traitement de la parole. Elles sont utilisées en traitement de la parole dans divers champs d'applications [6, 10, 11]. Par exemple, elles peuvent permettre d'extraire des annotations pertinentes de transcriptions automatiques ou encore d'améliorer l'efficacité des systèmes de dialogue grâce à la détection d'erreurs et des mots hors-vocabulaire. Les mesures de confiance sont généralement estimées grâce aux probabilités *a posteriori* données par le système de reconnaissance. Par exemple, dans [11], les auteurs ont recours aux graphes de mots et aux listes des N meilleures hypothèses pour définir la probabilité *a posteriori* d'un mot comme mesure de confiance. Un inconvénient de la probabilité *a posteriori* d'un mot est la forte sensibilité de cette mesure à la topologie de l'espace de recherche sur lequel elle est calculée (profondeur d'un graphe de mots ou taille d'une liste de N meilleures hypothèses). Ainsi, les différentes heuristiques de réduction de l'espace de recherche utilisées lors du processus de reconnaissance ont une incidence directe sur la valeur de la probabilité *a posteriori* et peuvent altérer sa pertinence.

Dans cet article, nous proposons d'améliorer les capacités de discrimination de la probabilité *a posteriori* en la combinant avec une mesure qui n'est pas affectée par ce type de problème. Cette nouvelle mesure est basée sur le comportement du repli du modèle de langage. La fusion des deux mesures semble réduire l'impact de l'espace de recherche sur la probabilité *a posteriori* et améliore sa capacité à discriminer une hypothèse correcte d'une hypothèse incorrecte.

2 Mesures de confiance

Soit l'ensemble de K mots reconnus $\{w_1, \ldots, w_K\}$. Chaque mot w peut être associé à une mesure de confiance $m(w)$, qui doit appartenir à l'intervalle $[0, 1]$ et doit correspondre à la probabilité que le mot w soit correct. Soit $\mu(m) = \frac{1}{K}\sum_{i=1}^{K} m(w_i)$, la dernière propriété nous donne que, pour une mesure de confiance idéale, $\mu(m)$ doit être une approximation de p_{ok} où p_{ok} est le taux de mots émis bien reconnus (par rapport au WER=$\frac{\#\text{sub}+\#\text{del}+\#\text{ins}}{\#\text{mots}}$, le taux de mots émis bien reconnus ne prend pas en compte les suppressions car il s'applique uniquement aux mots émis par le système de transcription).
Comparer la valeur prédictive $\mu(m)$ du taux de mots bien reconnus à la valeur réelle de ce taux peut être une bonne métrique pour évaluer la qualité d'une mesure de confiance. Mais cette métrique permet seulement l'évaluation de la capacité de prédiction globale de la mesure : elle ne reflète pas la pertinence locale de la mesure sur le mot. Cette information locale peut être évaluée grâce à une métrique utilisée lors des campagnes d'évaluation NIST. Cette métrique, appelée entropie croisée normalisée (NCE) sert à évaluer la pertinence des mesures de confiance, c'est une estimation de l'apport en information additionnelle d'une mesure de confiance [3] :

$$NCE = \{H_{max} + \sum_{correctw} \log_2(m(W)) + \sum_{incorrectW} log_2(1 - m(W))\}/H_{max} \quad (1)$$

où $H_{max} = -k\log_2(p_c) - (K-k)\log_2(1-p_c)$,
k est le nombre de mots reconnus corrects,
K , le nombre total de mots reconnus, et
p_c, la probabilité moyenne qu'un mot reconnu soit correct (= k/K).

Une autre métrique plus intuitive permet également d'apprécier les qualités d'une mesure de confiance. Il s'agit du CER (Confidence Error Rate) [10]. Cette mesure est simplement définie comme étant le nombre d'étiquettes incorrectement attribuées sur le nombre total de mots reconnus.
Une mesure de confiance doit donc respecter deux propriétés qui peuvent permettre de l'évaluer :
1. elle donne une prédiction globale sur la probabilté moyenne que le mot soit correct ;
2. elle donne une information locale sur la fiabilité d'un mot hypothèse.

Les mesures de confiance seront par la suite évaluées grâce à leur taux de prédiction globale ainsi qu'en termes de NCE et de CER.

2.1 Probabilité a posteriori

Les probabilités *a posteriori* peuvent être calculées à partir des listes des N meilleurs hypothèses [8], des graphes de mots [3] ou des réseaux de confusions [5]. En fait, la probabilité *a posteriori* d'un mot est le taux de la probabilité *a priori* d'un mot sur la somme des probabilités *a priori* de toutes les autres hypothèses alternatives. Ces probabilités *a priori* sont une combinaison des scores fournis par les modèles acoustiques et linguistique.
Dans les listes des N meilleurs hypothèses la probabilité *a posteriori* d'un mot est calculée avec le taux de la somme des probabilités *a priori* des occurrences de ce mot à une position donnée parmi les N hypothèses, sur la somme de toutes les probabilités *a priori* des mots situés à la même position, incluant celles des occurrences du mot courant.
Dans les aproches basées sur les graphes de mots ou les réseaux de confusions, la probabilité *a posteriori* est la généralisation de l'approche précédente où la segmentation en mots et la profondeur de l'espace de recherche sont mieux pris en considération.
Comme cette mesure est dépendante des heuristiques d'élaguage des graphes de mots générés lors du processus de reconnaissance, les scores de confiance obtenus grâce à celle-ci peuvent être biaisés. Pour remédier à cela, [3] utilise un mapping par arbre de décision pour transformer les probabilités *a posteriori* des mots en véritables scores de confiance.

Ici, nous utilisons un réseau de confusion basé sur la technique utilisée dans [5] pour calculer les probabilités *a posteriori* des mots. Pour éviter le biais dû aux heuristiques d'élagage, nous combinons la probabilité *a posteriori* avec une autre mesure de confiance qui n'est pas affectée par la taille de l'espace de recherche. La mesure de confiance basée sur les probabilités *a posteriori* des mots sera notée WP.

2.2 Mesure de confiance linguistique basée sur le comportement du repli

La mesure de confiance que nous introduisons ici est basée sur le comportement du mécanisme de repli d'un modèle de langage n-gramme, comme dans [9]. Nous appelons $LMBB$ (Language Model Back-off Behavior) cette méthode.
Pour un mot donné, il s'agit de prendre en compte l'ordre du n-gramme le plus élevé associé à ce mot. Par exemple, si la séquence de mots "... il est temps de ..." est reconnue en utilisant un modèle de langage quadrigramme et que le quadrigramme [il est temps de] a été observé dans le corpus d'apprentissage du modèle de langage, alors le mot 'de' sera associé à l'ordre 4. Par contre, si ce quadrigramme n'a pas été observé, mais que le trigramme [est temps de] l'a été, alors le mot 'de' sera associé à l'ordre 3. De même, ce mot pourrait être associé à l'ordre 2 ou à l'ordre 1 le cas échéant, et même à l'ordre 0 dans le cas peu courant où les mots hors-vocabulaire peuvent être traités.

Un phénomène bien connu en reconnaissance de la parole est la propagation des erreurs : lorsqu'un mot est mal reconnu, les mots qui l'entourent sont souvent affectés par des erreurs. Dès lors, il semble très intéressant d'intégrer dans la mesure de confiance linguistique d'un mot des informations concernant son voisinage. En supposant que le comportement du mécanisme de repli d'un modèle de langage est un bon indicateur de la fiabilité de ce modèle, nous proposons de prendre également en compte l'ordre associé aux deux mots voisins du mot visé (le voisin de gauche et celui de droite). Chaque mot reconnu est alors associé à trois valeurs d'ordre de n-gramme.
Afin de ne pas distinguer un nombre de classes différentes trop important qui seraient difficiles à bien modéliser sans grande quantité de données d'apprentissage, nous ne prendrons pas les valeurs réelles des ordres associés aux mots voisins mais leur position relative par rapport à l'ordre associé au mot visé : plus grand (+), plus petit (-) ou égal (=). Ceci permet de réduire le nombre de classes possibles.

Pour illustrer ce propos, prenons la séquence de mots "... il est temps de lire ce livre..." et supposons :
- que le quadrigramme [il est temps de] et le trigramme [est temps de] n'ont pas été observés dans le corpus d'apprentissage du modèle de langage, alors que le bigramme [temps de] l'a été : le mot 'de' est associé à l'ordre 2 ;
- que le quadrigramme [est temps de lire] n'a pas été observé dans le corpus d'apprentissage alors que le trigramme [temps de lire] l'a été : le mot 'lire' est associé à l'ordre 3 ;
- que le quadrigramme [temps de lire ce] a été observé dans le corpus d'apprentissage : le mot 'ce' est associé à l'ordre 4.
La classe de comportement du mot 'lire' sera alors associée à l'étiquette (-,3,+), car le mot 'lire' est associé à l'ordre 3, son voisin de gauche est associé à un ordre inférieur (-) de valeur 2 et son voisin de droite est associé à un ordre supérieur (+) de valeur 4.
Pour chaque classe et sur un ensemble de transcription donné, on peut calculer le taux d'erreur de reconnaissance des mots composant chacune de ces classes. Ce taux d'erreur est le rapport entre le nombre de mots $n_{err}(cl)$ mal reconnus (substitutions ou insertions) contenus dans une classe cl sur le nombre de mots $n_{mots}(cl)$ qui composent cette classe.Ainsi, pour un mot w associé à la classe cl, la valeur $m_{lmbb}(w)$ donnée par la mesure de confiance LMBB se calcule à partir d'un corpus d'apprentissage avec la formule :

$$m_{lmbb}(w) = 1 - \frac{n_{err}(cl)}{n_{mots}(cl)}$$

La figure 1 montre l'existence d'une corrélation entre comportement du mécanisme de repli du modèle de langage et taux d'erreur. Ces résultats ont été calculés sur le corpus d'apprentissage décrit section 3.2.

2.3 Fusion WP/LMBB

Fusionner la probabilité *a posteriori* avec une autre mesure qui n'est pas affectée par la taille de l'espace de recherche et qui est indépendante de celle-ci devrait améliorer ses résultats en tant que mesure de confiance. En effet, même si la mesure LMBB est dé-

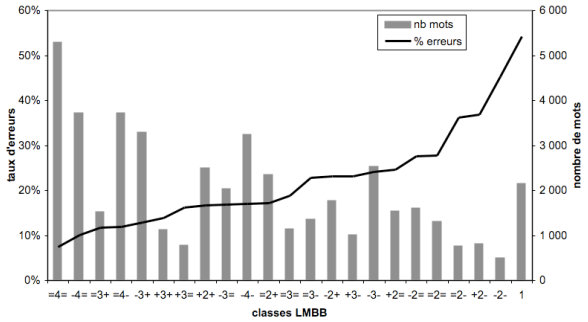


Fig. 1: Taux d'erreur, répartition des mots transcrits et classes LMBB

rivée du modèle de langage utilisé lors du décodage, ce n'est pas le score de celui-ci qui est pris en considération. Cette mesure apporte donc d'autres informations qui ne sont pas prises en compte lors du calcul de la probabilité *a posteriori*.

Combiner plusieurs paramètres pour obtenir une mesure unique peut être fait de plusieurs façons [7]. Les opérateurs les plus utilisés dans cette fusion sont : minimum, maximum, moyenne arithmétique, moyenne géométrique, produit ou encore la moyenne quadratique. Pour prendre en compte les qualités de chacune des mesures, une moyenne pondérée $m(w) = \sum_{j=1}^{J} q_j m_j(w)$, avec $\sum_{j=1}^{J} q_j = 1$ peut être utilisée. Les poids q_j peuvent être appris empiriquement par validation croisée (voir résultats de la section 2). Diverses techniques de fusion provenant de théories telles que la théorie de l'évidence ou encore la théorie des probabilités ont été essayées mais la technique qui a donné les meilleurs résultats sur le corpus d'apprentissage des mesures de confiance (CTrain) est une simple interpolation linéaire. La mesure retenue est notée WP/LMBB.

3 Expériences et résultats

Les différentes expériences décrites dans cet article sont basées sur le corpus d'ESTER, campagne d'évaluation sur des systèmes de transcriptions d'émissions radiophoniques en français démarrée en 2003 et achevée en janvier 2005 [4].
Le système utilisé durant cette campagne par le Laboratoire d'Informatique de l'Université du Maine (LIUM) est basé sur le décodeur CMU Sphinx 3.3. Plusieurs paramètres ont été ajoutés, comme l'adaptation de modèles acoustiques utilisant la méthode SAT (Speaker Adaptative Training) ou encore le rescoring de graphes de mots utilisant des modèles de language quadrigrammes. Ce système a atteint la seconde position de la campagne avec 23.6% de taux d'erreur mot (WER) –incluant les insertions, substitutions et suppressions [1, 2].

3.1 Modèles acoustiques et linguistique

Le dictionnaire utilisé contient environ 65K mots. Les modèles acoustiques ont été appris avec 81h de transcriptions manuelles provenant de différentes radios : France Inter, France Info, Radio Télévision Marocaine (RTM) et Radio France International (RFI). Ce sont des émissions d'actualités. Ces émissions sont majoritairement en bande large mais comportent également de de la bande étroite (téléphone). Pour cet apprentissage, nous avons utilisé le toolkit SphinxTrain, associé aux décodeurs de Sphinx CMU. Les modèles acoustiques sont appris avec une différenciation bande large/bande étroite. Le corpus de développement pour réévaluer les différents paramètres est constitué de 4h provenant de ces différentes radios. Le modèle de langage trigramme est utilisé lors des deux premières phases de décodage. La troisième phase utilise un modèle quadrigramme qui correspond à un rescoring de graphes de mots.

3.2 Apprentissage des paramètres pour les mesures de confiance

Les mesures de confiance ont été élaborées grâce à un échantillon de 4h provenant de France Inter, France Info, RTM et RFI. Ces 4h sont indépendantes du corpus d'apprentissage des modèles acoustiques et linguistique et sont fournies avec leur transcription manuelle. Une transcription automatique est obtenue avec le système de reconnaissance. Á partir des transcriptions automatique et manuelle, nous avons pu calculer les scores de confiance de la mesure LMBB à partir du taux d'erreur obtenu avec les différentes classes de LMBB (voir figure 1). Les probabilités *a posteriori* des mots sont calculées sur les graphes fournis par le système. Plusieurs mesures fusionnées ont été comparées sur CTrain et la meilleure en termes de NCE est : $m_{WP/LMBB}(w) = 0.7 * m_{WP}(w) + 0.3 * m_{LMBB}(w)$.

3.3 Résultats

Le corpus de test officiel (noté CTest) est composé de 10h d'émissions de radio (environ 10 000 phrases et 114 000 mots) : 2h de chacune des 4 radios du corpus d'apprentissage ainsi que 2h provenant de deux radios inconnues au moment de l'évaluation.
Nous avons vu à la section 2 que la moyenne des scores donnée par une mesure doit approximer le taux de mots émis bien reconnus. Le tableau 1 montre que la probabilité *a posteriori* ne permet pas d'obtenir ce taux sur les données de test. C'était aussi le cas sur CTrain. Pour pallier ce problème, une méthode classique de normalisation par transformation sigmoïdienne a été calculée mais cette méthode ne donnait pas d'amélioration en termes de NCE. Par contre, le taux de prédiction globale de la mesure LMBB est proche du taux de mots émis corrects. Ceci s'explique par le fait que les corpus CTest et CTrain sont proches en termes de couverture du modèle de langage. De plus, les classes LMBB ont été apprises sur CTrain et sont donc bien représentatives de l'impact du comportement du repli du modèle de langage. La mesure fusionnée WP/LMBB propose donc un taux de prédiction globale amélioré par rapport à celui de la probabilité *a posteriori* seule.
De plus, nous avons remarqué que beaucoup de mots ont une probabilité supérieure à 0,9 et pourtant, ils ne sont pas corrects. Une explication possible est que, dans l'espace de recherche, ces mots n'ont pas ou peu d'alternative. Néanmoins, ces mots ne sont pas pertinents du point de vue de la mesure LMBB et ont donc un score LMBB faible. La mesure LMBB permet donc à la probabilité *a posteriori* d'être plus discriminante.

Enfin, pour évaluer les mesure localement sur les deux corpus grâce à la NCE, le tableau 2 montre que la probabilité *a posteriori* donne une réelle information sur l'exactitude d'un mot d'une manière plus significative que la mesure LMBB. La mesure WP/LMBB obtenue en fusionnant les deux mesures améliore nettement les résultats de la probabilité *a posteriori* prise seule car les scores augmentent de 0,187 à 0,270 sur CTest. Ceci est dû au fait que les deux mesures initiales apportent des informations complémentaires.
Pour calculer le CER des différentes mesures, on a appliqué sur CTrain un seuillage des scores des mesures de confiance. Au-delà de ce seuil, les mots sont considérés comme étant corrects et en dessous, ils sont considérés comme incorrects. On peut alors vérifier ces hypothèses et déterminer le taux d'étiquettes incorrectement attribuées sur le nombre total de mots reconnus. Pour le seuil ayant obtenu le CER le plus faible sur CTrain, on calcule le CER sur CTest. Les resultats sont indiqués dans le tableau 3. Pour WP/LMBB, le seuil optimal est de 0,411. Pour la baseline, le CER est égal au nombre d'insertions et de substitutions divisé par le nombre total de mots reconnus. La fusion des mesures de confiance nous permet d'améliorer le CER par rapport à la probabilité *a posteriori* seule et nous constatons également un gain de 5% en relatif par rapport à la baseline.

Tab. 1: Taux de prédiction globale des mesures de confiance sur les données de test

Mesure	Prédiction globale
LMBB	83,3%
probabilité *a posteriori*	72,6%
WP/LMBB	75,2%
Taux correct réel	***80,8***%

Tab. 2: Comparaison des diverses mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes d'entropie croisée normalisée(NCE)

Mesure	CTrain	CTest
LMBB	0,081	0,063
probabilité *a posteriori*	0,169	0,187
WP/LMBB	**0,278**	**0,270**

Tab. 3: Comparaison des diverses mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes de Confidence Error Rate (CER)

Mesure	CTrain (%)	CTest (%)
LMBB	15,11	22,23
probabilité *a posteriori*	13,56	18,64
baseline	**15,09**	**19,23**
WP/LMBB	**13,17**	**18,27**

4 Conclusion

Dans cet article, nous proposons une amélioration de la probabilité *a posteriori* prise comme mesure de confiance. Pour obtenir une meilleure mesure, nous la fusionnons avec une autre mesure provenant d'une autre partie du système de reconnaissance pour bénéficier d'informations complémentaires à apporter à la probabilité *a posteriori*. Cette mesure est calculée à partir du comportement du repli du modèle de langage et est négligeable en termes de temps de calcul. Leur fusion, WP/LMBB, est une simple interpolation linéaire des deux mesures et améliore nettement leurs capacités à prédire si le mot émi est correct ou non. Cette mesure de confiance est pertinente pour de multiples applications du traitement de la parole. Avec une telle mesure, les mots fiables sont plus aisément détectables pour exploiter les sorties du système de reconnaissance, ou encore pour vérifier les capacités du système sur des tâches spécifiques. Par la suite, il serait intéressant d'essayer d'autres mesures pertinentes à fusionner ainsi que d'autres méthodes de fusion comme les arbres de décisions. Enfin, il serait intéressant d'étudier la mesure WP/LMBB dans des applications du traitement de la parole telle que le processus de décodage.

Références

[1] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[2] Y. Estève, P. Deléglise, and B. Jacob. Système de transcription automatique de la parole et logiciels libres. *Traitement Automatique Des Langues*, 45(2), 2004.

[3] G. Evermann and P.C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Speech Transcription Workshop*, 2000.

[4] G. Gravier, J.-F. Bonastre, S. Galliano, and E. Geoffrois. The ESTER evaluation campaign of rich transcription of french broadcast news. In *LREC, Language Evaluation and Resources Conference*, Lisbon, Portugal, May 2004.

[5] H. Mangu, E. Brill, and Stolcke A. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer Speech and Language*, pages 4373–400, 2000.

[6] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo. Confidence measures for spoken dialogue systems. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001.

[7] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 875–878, Munich, Allemagne, April 1997.

[8] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 163–166, Rhodes, Greece, 1997.

[9] C. Uhrick and W. Ward. Confidence metrics based on n-gram language model backoff behaviors. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Greece, September 1997.

[10] F. Wessel, K. Macherey, and R. Schlüter. Using word probabilities as confidence measures. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 225–228, Seattle, Washington, May 1998.

[11] F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13 :23–31, 2005.

Facteurs caractérisant les hésitations dans les grands corpus : langue, genre, style de parole et compétence linguistique

Ioana Vasilescu & Martine Adda-Decker

LIMSI-CNRS, Bât. 508, BP 133, 91403 Orsay cedex, France
Mél : ioana, madda@limsi.fr

ABSTRACT

This paper deals with the factors characterizing the autonomous vocalic filled pauses in large spontaneous speech corpora, namely language, gender, speaking style and language proficiency. Two corpora are analyzed: a corpus of broadcast news in French and American English and a corpus of short talks in a conference in English spoken by native and non-native speakers. Several acoustic and prosodic parameters are evaluated and correlated with each factor, namely timbre, pitch, duration and density. Results presented here show that the timbre is correlated with language and language proficiency, whereas the duration is linked both to gender and speaking style, the latter conditioning also the hesitation density in speech.

1. INTRODUCTION

Nous nous intéressons ici aux voyelles d'hésitation autonomes dans les grands corpus. Les hésitations représentent un des multiples phénomènes dits de « disfluence » recensés dans toutes les langues. Parmi eux notons l'allongement vocalique, les répétitions, les reformulations etc. Les hésitations vocaliques autonomes représentent un phénomène largement rencontré dans les langues qui consiste en l'insertion « à tout moment » dans le flux de parole spontanée d'un segment vocalique plutôt allongé. Ce segment vocalique peut être accompagné ou non de segments adjacents (coda nasal en anglais, diphtongaison etc.). Son rôle est « d'annoncer l'initiation de ce qui est attendu comme […] un délai dans la parole » (notre trad.) [1]. Les hésitations vocaliques n'ont pas de support lexical, ce qui les différencie de phénomènes similaires tels les allongements vocaliques d'un segment appartenant à un item lexical précis (généralement un mot grammatical).

La réalisation vocalique n'est pourtant pas la seule rencontrée parmi les langues du monde, d'autres phénomènes d'hésitations autonomes sont dénombrés, telles des consonnes nasales allongées (« mm » en chinois mandarin, par exemple), ou des demonstratives délexicalisés (« ano », « eto » en japonais, par exemple) [2,3]. Nous prenons en compte ici uniquement les hésitations vocaliques autonomes en français (« euh ») et en anglais (« uh », « um » en anglais américain / « er » en anglais du Royaume-Uni).

Antérieurement nous avons comparé les hésitations vocaliques autonomes en 8 langues : anglais américain, arabe, allemand, chinois mandarin, français, italien, espagnol sud-américain, portugais européen. Nous nous sommes intéressées à la voyelle support de chaque hésitation. Nous appelons *voyelle support*, la voyelle la plus longue et la plus stable de chaque occurrence. Cette voyelle est parfois diphtonguée à la fin ou suivie d'une coda nasale, comme en anglais. La voyelle support constitue en général l'élément principal d'une hésitation. Parmi les paramètres considérés (*durée*, *hauteur* et *timbre*) il s'est avéré que le *timbre* est le paramètre dépendant de la langue qui caractérise le mieux les hésitations. La *hauteur* et la *durée* permettent de différencier la voyelle d'hésitation des voyelles intra-lexicales de timbre similaire au sein d'une même langue. L'analyse inter-langue des paramètres hauteur et durée n'a pas révélé des différences majeures parmi les 8 langues considérées. Nos analyses ont confirmé des observations antérieures, i.e. l'hésitation vocalique autonome est significativement plus longue que les segments intra-lexicaux de timbre similaire et possède un contour F0 plat et stable [4]. Nous avons formulé l'hypothèse que le paramètre *timbre* est dépendant de la langue tandis que les paramètres *hauteur* et *durée* tendent à être des critères universels.

Nous considérons ici quatre facteurs susceptibles d'influer la production d'hésitations autonomes dans les grands corpus. Il s'agit des facteurs *langue* ; *genre* ; *style de parole* et *compétence linguistique* (langue maternelle, langue seconde).

2. CORPUS ET METHODOLOGIE

Deux corpus ont été utilisés dans cette étude : un corpus de journaux télévisés en anglais américain et en français et un corpus d'enregistrements d'interventions orales dans une conférence en anglais. Dans le corpus de journaux télévisés en anglais américain (désormais JTA) et le français (désormais JTF) sont parlés par des natifs, hommes et femmes. Nous comptons 6 sources en anglais américain (CNN, VOA, ABC etc.) et environ 150 locuteurs (dont 2 fois plus d'hommes) et 4 sources en français (France Inter, France Info, France2, France3) et environ 130 locuteurs (100 hommes, 30 femmes). Au total, environ 200 heures pour JTF et 100 heures pour JTE ont été utilisées. Le corpus d'interventions dans une conférence est « Terrible English Database » et consiste en des enregistrements de 10 minutes environ par des locuteurs natifs et non natifs d'anglais dans la conférence Eurospeech de 1993 [5]. Nous avons sélectionné 8 locuteurs français (désormais TEDF) et 3 locuteurs anglais (désormais TEDA). Ce sous-corpus préliminaire

ne contient pas d'intervenants femmes. Nous avons utilisé environ 1h30 d'enregistrements.

Les hésitations ont été automatiquement extraites et manuellement vérifiées afin d'éliminer des erreurs d'extraction potentielles ou des hésitations accompagnées de phénomènes non-verbaux pouvant jouer sur le calcul des paramètres (i.e. rires, bruits de bouche, bip de téléphone). Différemment de nos études précédentes, aucun critère de durée n'a été employé afin d'avoir une estimation réaliste de la durée et de la densité du phénomène par langue et par style de parole [6,7]. Le nombre d'occurrences par langue et corpus est montré ci-dessous.

Tableau 1 : Nombre d'hésitations pour les Français (h/f) et Anglais (h/f) parlant la langue maternelle (L1) ou une langue seconde (L2).

Corp./Loc.	**Français**	**Anglais**
JT	L1 : 1640 (h)/270 (f)	L1 :4455 (h)/491 (f)
TED	L2 :762 (h)	L1 :439 (h)

Les paramètres suivants ont été considérés : fréquence fondamentale (F0), timbre (F1/F2), durée et densité (durée totale hésitations/durée totale corpus).

3. LES FACTEURS *LANGUE* ET *GENRE*

Dans des études préliminaires antérieures, nous nous sommes intéressées aux particularités dépendantes de la langue *vs* universelles caractérisant les hésitations vocaliques autonomes. A cet effet, nous avons exploité des données de type « journaux télévisés » en huit langues. La présente étude prend en compte uniquement des hésitations en anglais et en français. Les données sont importantes quantitativement et permettent de rendre compte du phénomène d'hésitation à travers un nombre de paramètres susceptibles d'influencer ses caractéristiques acoustiques et prosodiques.

Les données analysées ici et issues des deux corpus de JT en français *vs* anglais américain confirment ces observations, notamment en ce qui concerne *le timbre*. Ainsi, la voyelle d'hésitation en anglais américain est significativement plus ouverte (F1) et plus antérieure (F2) que sa correspondante française (t-tests séries appariées, $p<0,0001$). Cette différence est indépendante de la variable genre (Figure 1).

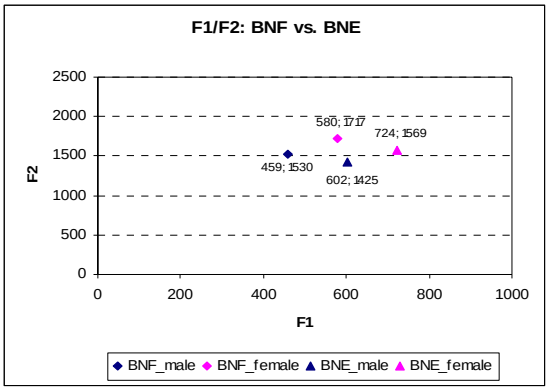


Figure 1 : Dispersion des valeurs moyennes des voyelles support dans un espace F1 *vs.* F2.

Le paramètre *hauteur* (F0) n'exhibe pas de différences notables entre les deux populations. En effet, la distribution des valeurs est équivalente, notamment pour les hésitations produites par les locuteurs hommes (Figure 2). En ce qui concerne les locutrices, il semblerait d'après la figure 2 que l'étendue pour les locutrices françaises est plus importante, avec plus de valeurs extrêmes. Notons toutefois que les données en français sont quantitativement moins importantes, ce résultat pouvant ainsi être un effet de corpus.

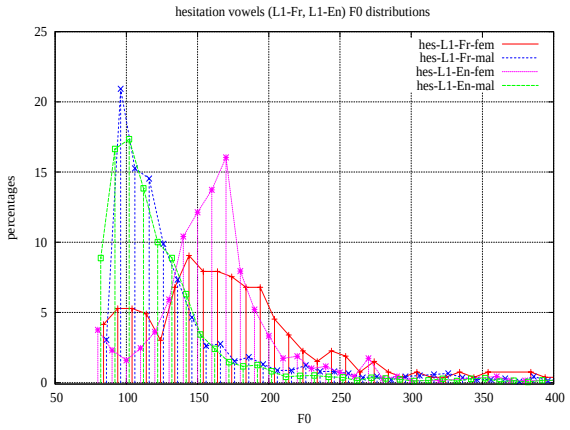


Figure 2 : Distribution des valeurs de F0 des voyelles support dans JT (anglais et français, hommes et femmes).

Le troisième paramètre analysé est la *durée*. Les données étudiées ici confirment la tendance observée dans d'autres langues, à savoir que la voyelle d'hésitation est significativement plus longue qu'une voyelle intra-lexicale. Elles mettent toutefois en évidence d'autres particularités (Tableau 2).

Tableau 2 : Durées moyennes des hésitations dans les corpus JTF et JTA (hommes et femmes)

Corpus/genre (ms)	**Hommes**	**Femmes**
JTF	343	262
JTA	267	266

Ainsi, en ce qui concerne la variable *langue*, il apparaît que les hésitations produites par les locuteurs hommes en français sont d'une durée significativement supérieures à celle des hésitations des locuteurs d'anglais américain (ANOVA, F=237,102, p<0,0001). Cette observation ne concerne pas les locutrices et, globalement, une différence significative entre les deux langues n'a pas été notée. Nous avançons l'hypothèse que la différence concernant les locuteurs serait due à un effet de corpus et non pas à un effet de langue.

Le dernier aspect considéré concerne la structure segmentale des hésitations. Alors qu'en français l'hésitation type est « euh », donc une voyelle centrale, l'anglais américain en possède deux. Comme observée plus haut, il s'agit d'une voyelle plus ouverte et plus antérieure que la correspondante française, suivie ou non d'un segment consonantique nasal. Le corpus JTE montre que les réalisations avec coda nasale sont minoritaires, à savoir 23% des hésitations seulement sont produites avec coda nasale. Plus encore, les réalisations avec coda nasale sont plus spécifiques aux femmes (45%) qu'aux hommes (19%).

4. LE FACTEUR *STYLE DE PAROLE*

Afin d'évaluer l'impact du facteur *style de parole* nous avons analysé et comparé les données de JT avec des données de TED. Deux conditions d'élocution sont ainsi mises en parallèle. Il s'agit d'une part du journalisme télévisé, impliquant une parole semi-préparée et des professionnels des interventions orales dans un temps limité, et d'autre part d'orateurs présentant leurs travaux à un public-juge. Ces derniers sont plus susceptibles de subir l'effet du stress, d'autant plus que pour 8 d'entre eux l'intervention se fait dans L2. Cette partie de l'étude prend en compte les productions des hommes dans le corpus JT, le corpus TED décrit ici comportant pour l'heure les productions à 11 locuteurs.

La *densité* est un paramètre qui s'avère dépendant du facteur *style de parole*. La durée totale des hésitations représente 0,7% du corpus JTA et 0,1% de JTF, tandis que dans TEDA il s'agit de 5,8% et dans TEDF de 5,7%. Cette différence semble conforter l'hypothèse que des facteurs tels que le stress, la parole non-planifiée (absence des prompteurs aidant les locuteurs des corpus JT) et le fait que les locuteurs soient non-professionnels, pourraient jouer sur le pourcentage de disfluences présentes dans le discours. Le statut de langue maternelle vs. seconde langue ne semble pas avoir influencé le paramètre *densité* : elle est comparable dans le corpus TEDF vs. TEDA.

En ce qui concerne la *hauteur* F0 nous n'avons pas noté des différences significatives entre les valeurs moyennes dans les deux corpus.

Tableau 3 : F0 moyenne des hésitations dans les corpus JT et TED (hommes).

Corpus/F0_Moy (Hz)	**F0_Moyen**	**Ecart-type**
JTF	142	86
JTA	129	71
TEDF	129	42
TEDA	141	67

Le corpus TED montre une variabilité inter-locuteurs importante. Une analyse ANOVA prenant en compte les données des locuteurs français et anglais respectivement met en évidence un effet de locuteur statistiquement significatif (TEDF : ANOVA, F=9,7111, p<0,0001 ; TEDA : ANOVA, F=23,151, p<0,0001). Par ailleurs, il semble que les locuteurs français s'exprimant en anglais (L2) présentent une étendue plus importante des valeurs de F0, notamment en ce qui concerne le registre aigu (Figure 3). L'hypothèse pourrait être formulée qu'à la variabilité liée aux contraintes temporelles de la prise de parole en public s'ajoute celle de l'expression dans une L2. Ces facteurs pourraient influencer le contrôle du flux de parole et notamment de l'usage d'un registre modal. Cependant, plus de données seraient nécessaires pour valider cette hypothèse.

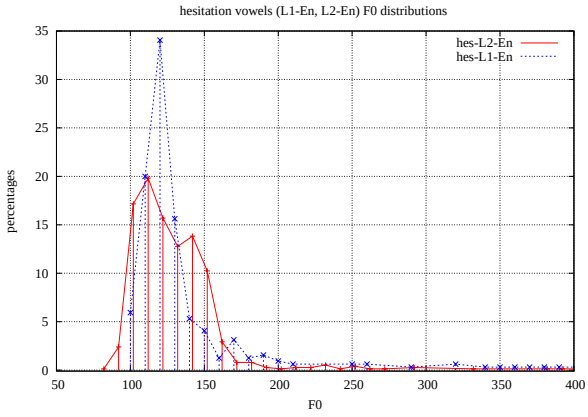


Figure 3 : Distribution des valeurs de F0 des voyelles support des hésitations dans le corpus TED.

La *durée* s'avère un paramètre sensible au facteur *style de parole*. Ainsi, les hésitations dans le corpus TED sont significativement plus longues que dans les corpus JT, et cela peu importe la langue (ANOVA, F=268,897, p<0,0001) (Tableau 3).

Tableau 4 : Durée moyenne des hésitations dans les corpus JT et TED (hommes).

Corpus/Durée moy. (ms)	**Durée moyenne**
JTF	342
JTE	266
TEDA	429
TEDF	415

Nous posons ici l'hypothèse que la différence de durée est liée au rôle des hésitations dans un corpus de type « journaux télévisés » *vs.* « conférence ». Dans une intervention de type conférence et, pour la plupart des locuteurs, menée dans une langue seconde, les hésitations marquent une vraie recherche du discours dans des conditions de stress. Dans les journaux télévisés les interventions sont souvent préparées et les hésitations pourraient avoir un caractère plus « canonique » que dans un contexte où la parole est moins contrôlée.

4. LE FACTEUR *COMPETENCE LINGUISTIQUE*

Nous appelons compétence linguistique le degré de maîtrise de l'anglais (L2) par les locuteurs français de TED. Ce facteur est évalué à travers le *timbre* des hésitations produites par les Français en anglais L2, par rapport au français L1, et à l'anglais L1.

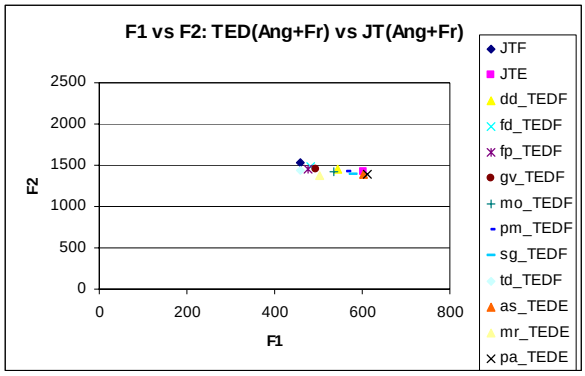


Figure 4 : Dispersion des voyelles support des locuteurs de JT et TED (JTF, JTA : moy./corpus ; dd, fd, fp, gv, mo, pm, sg, td : moy./loc. TEDF ; as, mr, pa : moy./loc. TEDA).

La figure 4 montre que les voyelles support en anglais (L2) se placent sur un continuum sur l'axe ouvert/fermé entre les valeurs des voyelles support en français et en anglais comme L1. Certains locuteurs français de TED produisent les hésitations de leur langue maternelle lorsqu'ils s'expriment en L2, tandis que d'autres produisent des voyelles support intermédiaires en termes d'ouverture, entre les valeurs de JTF et JTA.

5. DISCUSSION

Dans cette étude nous avons analysé plusieurs facteurs caractérisant les hésitations autonomes dans les grands corpus oraux. Les facteurs langue, genre, style de parole et compétence linguistique ont été évalués à travers les paramètres acoustiques et prosodiques traditionnellement mesurés pour décrire les hésitations vocaliques autonomes, i.e. timbre, durée, hauteur et densité. Le facteur *langue* a confirmé des observations antérieures, à savoir que le paramètre le plus distinctif est le timbre. La durée a été mise en lien à la fois avec le facteur *genre* et *style de parole*. Ce dernier doit être de plus mis en relation avec le paramètre densité. Enfin, la *compétence linguistique* se traduit par des productions intermédiaires en termes de timbre des voyelles support des hésitations en anglais (L2) par rapport au français et à l'anglais comme L1. Cette observation soulève des questions intéressantes liées au statut des disfluences lors de l'acquisition de L2.

BIBLIOGRAPHIE

[1] Clark H.H., Fox Tree J.E. 2002. Using uh and um in spontaneous speaking, *Cognition* 84, 73-111.

[2] Zhao, Y., Jurafsky, D., 2005, A preliminary study of Mandarin filled pauses, DISS05, Aix-en-Provence.

[3] Watanabe, M., Den, Y., Hirose, K. , Minematsu, N. (2005): "The effects of filled pauses on native and non-native listeners² speech processing", In *DiSS-2005*, 169-172.

[4] Shriberg, E., The 'errrr' is human: ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, 31/1, 2001.

[5] Lamel, L., Schiel, F., Fourcin, A., Mariani, J., Tillmann, H., "The translanguage english database ted", ICSLP 1994, Yokohama, Japon.

[6] Vasilescu, I. Candea, M., Adda-Decker, M., Hésitations autonomes dans 8 langues : une étude acoustique et perceptive, Workshop MIDL04, Paris France, 2004.

[7] Candea, M., Vasilescu, I., Adda-Decker, M., Inter- and intra-language acoustic analysis of autonomous fillers, DISS05, Aix-en-Provence, France.

Etude de dysfluences dans un corpus linguistiquement contraint

Jean-Leon Bouraoui, Nadine Vigouroux

IRIT,
Université Paul Sabatier
118, route de Narbonne,
31062 Toulouse, France
{bouraoui,vigourou}@irit.fr

ABSTRACT

This paper presents a study carried out on an air traffic control corpus which presents some specificity: apprenticeship situation, and the fact that the production is subordinate to a particular phraseology.

Our study is related to the many kinds of disfluency phenomena that occur in this corpus, and the way they are or not affected by the nature of the corpus. We define 6 main categories of these phenomena. We then present the distribution of these categories. It appears that some of the occurrences frequencies largely differ from those observed in other studies. Our explanation is based on the corpus specificity: in reason of their responsibilities, both controllers and pseudo-pilots have to be especially careful to the mistakes they could do, since they could lead to some dramas.

1. INTRODUCTION

Les dysfluences sont un phénomène apparaissant fréquemment dans toute production orale spontanée. Elles ont donné lieu à de nombreuses études, que ce soit dans le domaine du Traitement Automatique de la Parole, ou celui du Traitement Automatique du Langage Naturel. En effet, leur étude et leur identification précise est primordiale. Sur le plan théorique, pour mieux comprendre et modéliser les problèmes pouvant advenir lors de toute communication orale. Sur le plan applicatif également, les intérêts sont nombreux : par exemple, pour améliorer la robustesse des systèmes automatiques de dialogue oral, ou « nettoyer » automatiquement des transcriptions de productions orales spontanées (Adda-Decker *et al.* [1]).

Cependant, la majorité des études faites sur le sujet portent plus ou moins sur un langage quotidien : dialogues « à bâtons rompus », demandes d'horaire, discours, etc. Par contre, aucune étude n'a à notre connaissance été menée sur les dysfluences apparaissant dans un corpus d'oral spontané produit dans le cadre d'une tâche particulièrement contrainte, notamment d'un point de vue linguistique, par l'utilisation d'une phraséologie spécifique. Or, on sait que l'utilisation d'une « langue de spécialité »[1] dans le cadre d'une tâche donnée entraîne des comportements spécifiques. Ceux-ci touchent les plans linguistique et cognitif (cf. notamment Lerat [12] et Falzon [8]). On peut alors se demander si on peut également observer ce type de modifications concernant l'apparition des dysfluences dans l'oral spontané.

Répondre à cette question, ou du moins apporter des éléments précis d'information, est précisément le but de l'étude relatée dans le présent article. Elle porte en effet sur un corpus d'oral spontané consistant en dialogues relatifs au contrôle aérien, et devant respecter une stricte phraséologie.

La présentation de notre étude se fait en trois temps. Le premier, assez succinct, est consacré à la description du corpus et de ses caractéristiques. Dans un deuxième temps, nous donnons les différents types de dysfluences relevées dans le corpus, ainsi que leur distribution. Enfin, nous nous intéressons à un phénomène particulier de dysfluence, et à la manière dont celui-ci est affecté par la nature du dialogue et de la tâche.

2. PRÉSENTATION DU CORPUS

2.1. Les communications entre contrôleurs et pseudo-pilotes

Le corpus sur lequel porte notre étude est constitué d'enregistrements de dialogues oraux spontanés entre des contrôleurs aériens en formation et des « pseudos-pilotes ». Ces derniers sont des instructeurs simulant le rôle de pilotes en exercice. Deux langues sont utilisées : le français (majoritaire) et l'anglais. Le but des exercices enregistrés est d'entraîner les apprentis contrôleurs, et ensuite de les évaluer. Il s'agit pour eux de gérer plusieurs avions situés dans la zone contrôlée, par exemple en leur assignant une vitesse et/ou une position données. Pour des raisons techniques, le canal audio ne peut être « occupé » que par un seul locuteur à la fois, ce qui empêche tout recouvrement de parole.

Les productions orales des contrôleurs et des pilotes sont gouvernées par une stricte phraséologie, présentée dans [2]. Celle-ci décrit, par exemple, la manière dont les locuteurs doivent prononcer les identifiants des avions, ou bien l'ordre que doivent observer les différentes parties d'un message[2]. Durant la formation, ainsi d'ailleurs que dans des conditions réelles de travail, la phraséologie n'est pas toujours systématiquement appliquée. Le cadre général qu'elle fixe est cependant respecté.

[1] Plusieurs autres termes synonymes sont utilisés dans la littérature.

[2] Pour une description des indicatifs français et des ordres, se référer à Dourmap & Truillet [7].

Il est également important de noter que les dialogues enregistrés appartiennent bien à la catégorie du discours oral spontané. Nous tenons à le préciser car le rôle prépondérant tenu par la phraséologie pourrait laisser à penser que tous les énoncés produits sont déjà planifiés à l'avance. Or, ce n'est pas le cas : ni les contrôleurs, ni les pilotes ne savent à l'avance ce qui va arriver, et par conséquent ce qu'ils vont avoir à dire. La phraséologie définit seulement le cadre général de production des énoncés ; ce qui est dit repose sur l'interaction dynamique entre un contrôleur et pilote ou pseudo-pilote donnés, en fonction d'une situation variable.

2.2. Méthodologie de transcription et d'annotation

Nous avons procédé à la transcription et l'annotation des dialogues selon les spécifications de Coullon & Graglia [5].

Ces spécifications ont pour but de déterminer les éléments à transcrire, d'obtenir l'homogénéité des transcriptions dans le cas où plusieurs annotateurs se succèdent. Elles consistent essentiellement en règles à suivre pour transcrire les termes techniques tels que les indicatifs, les vitesses, etc. Elles donnent également des instructions de transcription des phénomènes tels que les hésitations ou les pauses. Nous avons ajouté à ces spécifications quelques autres classes et sous-classes de phénomènes devant être transcrits.

Le logiciel Transcriber[3] 1.4.2 a été utilisé pour les transcriptions.

2.3 Caractéristiques du corpus

Les enregistrements ont été effectués avec un DAT (Digital Audio Tape), et échantillonnés à 16 kHz (16 bits). Pour des raisons d'enregistrement, la qualité sonore souffre parfois de problèmes résultants de la saturation ou de bruits tels que les interférences ; cependant, les dialogues sont intelligibles. La table 1 ci-dessous présente les principales caractéristiques du corpus[4].

Table 1: Principales caractéristiques de notre corpus

Durée	Nombre de locuteurs	Nombre de tours de parole	Nombre de mots
36h50mn	16 (répartis en 2 groupes)	11 427	76 306

3. LES PHÉNOMÈNES DE DYSFLUENCE

3.1. Quelques points de terminologie

Dans la littérature, les termes utilisés par les auteurs pour désigner un phénomène donné varient souvent. Pour cette raison, nous présentons ci-dessous la terminologie (en français) employée dans notre travail.

Nous avons défini 6 différentes catégories de dysfluences. Lorsque cela est nécessaire, nous donnons des exemples pour illustrer notre propos (en mettant la dysfluence en gras).

- Hésitation : désigne l'interjection « euh ». Selon certaines terminologies (notamment Henry *et al.* [10]), il appartient à la catégorie des « pauses remplies ». Exemple:

*maintenons niveau 1 0 0 Poitiers Amboise **euh** Lacan*

- Répétitions : un mot (ou un groupe de mots) apparaît au moins deux fois à la suite. Nous n'avons pas pris en compte la répétition de dysfluences. Exemple:

*station **station** calling euh repeat your callsign*

- Amorce : l'arrêt de la production d'un mot avant la fin normale de celui-ci. Dans notre terminologie, une amorce correspond toujours à un fragment de mot que l'on peut identifier (souvent grâce à la connaissance de la phraséologie). Exemple (l'amorce est entre crochets) :

*speed euh 200 Kts **[mak]** euh minimum*

Le contexte et la phraséologie aident à comprendre que le contrôleur commence à prononcer « maximum ». Il se rend compte que cela ne convient et s'interrompt (« mak »). Enfin, il dit le mot correct : « minimum ».

- Fragment de mot : un ou plusieurs phonèmes inidentifiables (par opposition aux amorces). Exemple (le fragment est entre crochets) :

*due to **[ou]** due traffic euh descend level 9 0*

- Allongement : l'allongement d'une unité phonétique d'un mot, supérieur à 0,5 sec. Peut être combiné aux hésitations. Ce phénomène entre également dans la catégorie des « pauses remplies ».
- Pause longue : toute pause supérieure à 0,5 seconde et comprise à l'intérieur un tour de parole

3.2. Distribution des phénomènes

La figure 1 présente la répartition des catégories décrites ci-dessus. Les nombres situés immédiatement après le nom de la catégorie correspondent au nombre total d'occurrences relevées ; les pourcentages (en gras) sont calculés par rapport au nombre total d'occurrences de dysfluences.

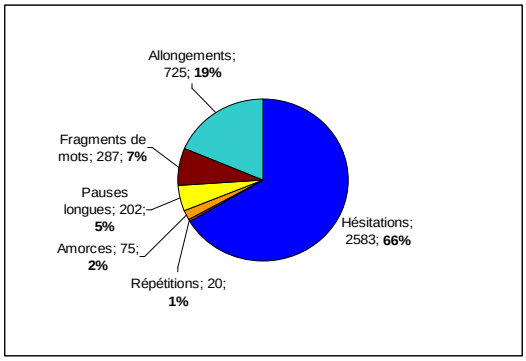


Figure 1: Distribution des dysfluences

[3] http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html

[4] Le corpus (oral et transcription) n'est pas disponible sans demande préalable auprès de l'ENAC. S'adresser aux auteurs pour plus de renseignements, ou pour demander des échantillons.

Mettons cette distribution en perspective au moyen d'une comparaison détaillée avec d'autres études sur le même thème. Voici d'abord une courte description (nature de la tâche, nombre de mots, etc.) de chaque corpus d'oral spontané sur lesquels ces études sont basées :

- Candéa [4][5] : corpus de 13 histoires racontées oralement par des enfants. Durée : 70 minutes et 25 secondes ;
- Henry & Pallaud [9] : corpus de 1 000 382 (différentes situations d'oral spontané ; 794 locuteurs différents);
- Shriberg [13] : corpus de 54 minutes, et comprenant 8500 mots. Les 10 locuteurs parlent de leur travail ou de leur souvenirs ;
- Kurdi [11] : basée sur corpus de négociations (en Anglais) de transport de marchandises par train. Il comprend 52 000 mots.

On le voit, ces études sont basées sur des corpus très variés, que ce soit au niveau de la tâche ou de la taille. Cette diversité constitue une base pertinente de comparaison avec notre propre corpus.

Nous allons maintenant présenter les comparaisons pour chacune des catégories de dysfluences que nous avons définies. Evidemment, puisque chaque étude ne couvre pas l'ensemble des dysfluences, nous présenterons seulement celles qui concernent un phénomène donné, ou dont la catégorisation est proche de la nôtre[6].

- Répétitions : on voit par la table 2 que notre corpus comprend beaucoup moins de répétitions que les autres corpus. Nous pensons que la principale explication repose sur la nature même du corpus. Notre hypothèse est que, dans un contexte de contrôle aérien, les locuteurs (contrôleurs et pilotes) doivent être particulièrement vigilants afin d'éviter les ambiguïtés ou problèmes qui pourraient nuire à la compréhension de l'énoncé. De même, le temps nécessaire pour produire un énoncé n'est pas extensible : le locuteur ne peut pas perdre trop de temps en hésitation ou autres pauses (remplies ou silencieuses). Comme nous le verrons plus bas, cette hypothèse est confirmée par le fait que, pour chacune des autres catégories de dysfluences que nous avons définies, il y a systématiquement moins d'occurrences dans notre corpus (proportionnellement à la taille du corpus de comparaison) ;

Table 2: comparaisons pour les répétitions

Nom de l'étude	Notre corpus	[4]	[13]	[11]
Nombre de répétitions	20	110	141	256

- Hésitations : comme on le voit dans la table 3, il y a également bien moins d'hésitations dans notre corpus que dans ceux des autres travaux référencés. Ainsi, il y a 544 occurrences dans [4], mais ce corpus ne dure que 70 minutes, contre 35 heures pour le nôtre. De ce fait, il y a proportionnellement moins d'occurrences dans le corpus de [4]. On peut cependant noter que la différence semble globalement moindre que celle que nous avons observée pour les répétitions ;

Table 3: comparaisons pour les hésitations

Nom de l'étude	Notre corpus	[4]	[11]
Nombre et/ou pourcentage d'hésitations (par rapport au nombre total de mots)	2583 / **3.38%**	544	3512 / **6.75%**

- Amorces et fragments de mots : [9] est la seule étude dont la catégorisation est la plus proche de la nôtre en ce qui concerne les « amorces » et « fragments de mots ». Elle présente également des statistiques détailles sur leur distribution. Comme les auteurs ne font pas la distinction entre les « amorces » et les « fragments de mots », nous additionnerons les occurrences des deux types de phénomènes qui apparaissent dans notre corpus. Il en résulte un total de 362 occurrences, soit 0.47% du nombre total de mots. [9] fait état d'un total de 6094 occurrences des « fragments de mots » pour environ un million de mots (soit approximativement 0.6%). La distribution dans notre corpus de cette double catégorie est assez proche de celle observée dans [9], contrairement à ce que nous avons constaté pour les catégories précédentes. Toutefois, une explication à ce résultat pourrait être le fait que notre double catégorie ne correspond pas exactement à celle définie par [9] ;

Table 4: comparaisons pour les allongements

Nom de l'étude	Notre corpus	[4]	[13]
Nombre et/ou pourcentage d'allongements (par rapport au nombre total de mots)	725 / **0.9%**	284	669 (y compris "euh") / **7.9%**

- Allongements : une fois encore, comme on le voit dans la table 4, la fréquence de ce que nous appelons allongement est moindre dans notre corpus que dans les autres ;
- Pauses longues : la table 5 montre que la spécificité de notre corpus est un peu moins prononcée que pour les autres catégories de dysfluences. Cependant, là encore, on remarquera qu'il y a moins de " pauses longues" que dans d'autres corpus.

Table 5: comparaisons pour les pauses longues

Nom de l'étude	Notre corpus	[4]	[13]
Nombre et/ou pourcentage of pauses longues (par rapport au nombre total de mots)	827 / **1.08%**	1471	318 / **3.74%**

4. LE CAS DES AMORCES

Plusieurs intérêts scientifiques nous poussent à nous pencher sur le cas des amorces, bien qu'elles soient peu représentées dans notre corpus. Le principal est qu'elles permettent, dans de nombreux cas, d'avoir une idée de la planification de la production du locuteur, comme nous allons le montrer ci-après.

Plus particulièrement, nous nous sommes intéressés aux amorces correspondant à une auto-correction du locuteur : celles où l'arrêt en cours de production caractéristique des

[5] Pour plus de lisibilité, nous désignons dans la suite de l'article les études uniquement par leur numéro de référence.

[6] Lorsque cela est possible, nous indiquons dans les tableaux le nombre d'occurrences et le pourcentage. Nous mettons en gras cette dernière mesure, afin de faciliter la lecture.

amorces est motivé par la prise de conscience du locuteur qu'il fait une erreur[7]. Nous avons distingué quatre sous-catégories, en fonction précisément de la nature de l'erreur qui provoque l'interruption de la production :

- Erreur sur un « mot » : nous appelons « mot » les données alpha-numériques (indicatifs, par exemple), et les commandes (« grimpez », « demande », etc.) Exemple :

climbing for level 1 7 0 and 2 0 0 Kts ***[mak]*** *euh minimum D M C*

- Erreur sur l'organisation de l'énoncé : un mot ou un groupe de mots n'occupant pas sa position correcte dans l'énoncé.

[poi] *Absie Poitiers Balon Reson Britair B X*

- Erreur sur la langue utilisée : le locuteur remarque (ou bien on lui fait remarquer) qu'il n'a pas parlé dans la langue appropriée : français à la place de l'anglais ou vice versa.

P I ***[vite]*** *speed 2 1 0 Kts*

- Erreur de prononciation : comme son nom l'indique…

c'est le ***[lio]*** *Littoral*

La distribution de ces différentes sous-catégories est donnée dans la figure 2 ci-dessous.

On constate que la majorité des erreurs concerne ce que nous avons appelé « mot ». Nous attribuons cela à la charge cognitive élevée que cette catégorie peut induire. En effet, il s'agit très souvent de termes que les apprentis contrôleurs ne sont pas encore habitués à « manier ». La forte charge cognitive ainsi engendrée est elle-même la source de problèmes de production.

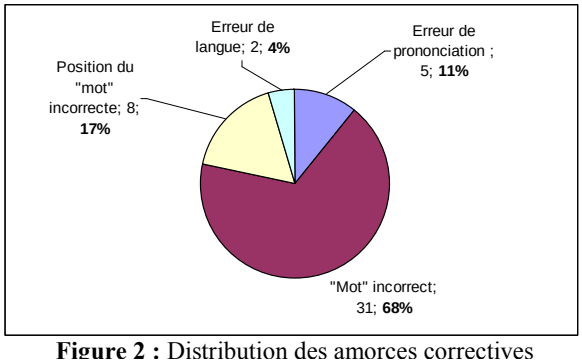


Figure 2 : Distribution des amorces correctives

5. CONCLUSION

Notre corpus présente un certain nombre de caractéristiques différentes de celles de corpus plus « traditionnels », par exemple de demandes d'informations. Nous avons montré qu'il y a une très forte différence entre les fréquences d'apparition de dysfluences dans notre corpus par rapport à d'autres corpus, notamment en ce qui concerne les répétitions. L'explication plausible de ce phénomène repose sur la spécificité de ce corpus. En nous penchant sur le cas plus particulier des « amorces », nous avons également mis en évidence la distribution particulière de cet autre type de phénomènes, que nous avons attribuée à la spécificité de la tâche.

De nombreuses prolongations à cette première étude sont envisagées, telles que la caractérisation acoustique (intensité, durée, fréquence fondamentale, représentation spectrale) des dysfluences, pour améliorer leur reconnaissance au niveau du décodage acoustique.

Une caractérisation plus fine des propriétés linguistiques du corpus dans la perspective d'une modélisation stochastique dans le cadre d'un module de compréhension de la parole est en cours. Elle nous permettra de procéder à l'implémentation du modèle obtenu.

Enfin, nous poursuivons un autre axe de recherche : l'analyse des rapports entre la situation d'apprentissage, la charge cognitive des locuteurs, et les performances de ceux-ci.

Nous mettons actuellement en place en place un protocole d'évaluation reproduisant des conditions de contrôle aérien aussi proche que possible de la réalité. Il nous servira à mettre en œuvre les deux dernières perspectives.

[7] Cela concerne 29 amorces, soit 39% de leur nombre total.

BIBLIOGRAPHIE

[1] M. Adda-Decker, B. Habert, C. Barras, G. Adda, P. Boula De Mareuil, P. Paroubek. Une étude des dysfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. *(JEP'04)*. 2004.

[2] Arrêté du 27 juin 2000 relatif aux procédures de radiotéléphonie à l'usage de la circulation aérienne générale. *J.O n° 171 du 26 juillet 2000*, p. 11501.

[4] M. Candéa. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse d'État, Université Paris III (Sorbonne Nouvelle), 2000.

[5] I. Coullon & L. Graglia *Spécifications de la base de données pour l'analyse des communications VHF en route.* CENA internal report, 2000.

[7] L. Dourmap & P. Truillet. *Interaction vocale dans le contrôle aérien : la comparaison de deux grammaires contextuelles pour la reconnaissance des indicatifs de vol.* CENA internal report. 2003.

[8] P. Falzon *Ergonomie cognitive du dialogue*. Presses Universitaires de Grenoble, 1989.

[9] S. Henry & B. Pallaud. Word fragments and repeats in spontaneous spoken French, *DiSS'03,* 2003.

[10] S. Henry, E. Campione & J. Véronis. Répétitions et pauses (silencieuses et remplies) en français spontané. *(JEP'04)*. 2004.

[11] M.- Z. Kurdi. *Contribution à l'analyse du langage oral spontané.* Thèse de doctorat, Université J. Fourier, Grenoble, France, 2003.

[12] P. Lerat. *Les langues spécialisées.* Paris ,PUF, 1995.

[13] E. Shriberg. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of Berkeley, California, 1994.

Intelligibilité de la parole après glossectomie totale et réhabilitation orthophonique précoce

Florence Fauvet[1&2], Philippe Schultz[1], Christian Debry[1], Fabrice Hirsch[2], Rudolph Sock[2]

[1]Service O.R.L. - Hôpitaux Universitaires de Strasbourg
1 av. Molière – 67098 Strasbourg – Cédex, France.
[2] E.A. 1339-Linguistique, Langues et Parole (LiLPa) – Composante Parole et Cognition
Institut de Phonétique de Strasbourg – Université Marc Bloch
22 rue Descartes – 67084 Strasbourg– Cédex, France.
Mél : florence.fauvet@chru-strasbourg.fr

ABSTRACT

This paper reports the monitoring and treatment of speech and its intelligibility following total glossectomy. A very early speech therapy rehabilitation in the post-operative course was initiated in a complex case of oral cavity surgery. Training was directly based on respiration, on mobility of articulators and on variations of the vocal tract volume for future production of consonants and vowels. The patient quickly recovered sufficient phonation capabilities to communicate with his relatives and medical staff. Although speech therapy and self-training ended progressively after the patient was discharged from hospital, speech was registered at different periods after rehabilitation. Spontaneous speech evaluation shows improvement of speech intelligibility and pronunciation. In such a case, accurate phonetics data have opened new therapeutic perspectives.

1. INTRODUCTION

Peu de données dans la littérature concernent la réhabilitation de la parole et son intelligibilité en cas d'exérèse totale de la langue [1] à [8]. Un programme de suivi orthophonique, quasiment limité à la période d'hospitalisation, a débuté quelques jours après intervention chez l'un de nos patients atteint d'un cancer de la cavité endobuccale. La chirurgie reconstructrice réalisée ayant pour conséquence l'immobilité de la "nouvelle langue", nous détaillons des procédés de réhabilitation précoce et des stratégies d'adaptation pour la reprise de la phonation. À partir d'enregistrements effectués à un mois et demi, deux puis cinq mois après opération, un jury d'écoute a évalué des échantillons de conversation spontanée du point de vue de son intelligibilité, de son débit, de sa prosodie et de son articulation. Nous présentons les résultats obtenus dans ce cas particulier de traitement thérapeutique, les possibilités de communication du patient et les perspectives éventuelles de prise en charge.

1. CAS CLINIQUE

Il s'agit d'un patient de 52 ans présentant une volumineuse tumeur pelvilinguale classée T4 N2c M0 fixant toute la langue et envahissant la symphyse mandibulaire. La prise en charge chirurgicale consiste en une pelviglossectomie avec mandibulectomie interruptrice emportant la totalité du plancher de bouche, la symphyse mandibulaire et la langue. Les sections chirurgicales de la base de langue se situent au niveau des sillons amygdaloglosses et des vallécules qui sont respectées. On préserve les nerfs X, XI, XII à droite et à gauche. La reconstruction mandibulaire repose sur la mise en place d'une prothèse en titane fritté fixée à chaque branche horizontale. Le larynx est libéré de ses attaches inférieures et suspendu aux branches horizontales de la mandibule et de la prothèse. Un lambeau musculo-cutané de grand pectoral recouvrant la prothèse est suturé entre le vestibule buccal inférieur et les vallécules linguales. Une trachéotomie de sécurité assure la ventilation du patient avant cicatrisation complète et une gastrostomie d'alimentation aura préalablement été posée.

La cicatrisation des voies d'abord et du lambeau est rapide. L'examen au nasofibroscope réalisé au dixième jour retrouve une stase salivaire au niveau des sinus piriformes. La mobilité du larynx est normale. L'alimentation par gastrostomie est maintenue en raison de l'inflammation pharyngée induite par la radiothérapie. La trachéotomie est elle aussi laissée en place afin de s'assurer de la liberté des voies aériennes, le patient pouvant s'exprimer à l'aide d'une canule de trachéotomie fenêtrée. Il devait cependant décéder 6 mois après l'intervention chirurgicale des suites de métastases pulmonaires.

3. MÉTHODE ET PROCÉDÉS DE RÉHABILITATION

3.1 Modalités de la prise en charge rééducative

Le suivi orthophonique a débuté quatre jours après l'intervention, sous couvert médical, dans le cadre de l'hospitalisation au service O.R.L. de l'Hôpital de Hautepierre (Hôpitaux Universitaires de Strasbourg). Onze séances au total ont été effectuées dont neuf sur une période de dix-sept jours et deux aux consultations de contrôle à environ deux mois de l'intervention. Leur durée, selon les circonstances et la fatigabilité du sujet, s'est inscrite de quelques minutes à une demi-heure. Le programme de réhabilitation établi a comporté des exercices progressivement plus nombreux. Le patient, qui s'est montré d'emblée motivé et heureux de pouvoir s'exprimer, s'est entraîné quotidiennement, à raison de trois à quatre fois par jour, trois fois de suite pour chaque exercice, en dehors des séances de rééducation, y compris au cours du mois qui a suivi son retour à domicile. Puis son état de santé s'est altéré

progressivement.

3.2 Objectif de la prise en charge

D'un point de vue général et dans ce cadre de pathologie, l'intervention de l'orthophoniste a pour objectif la réhabilitation de l'ensemble des fonctions de la cavité buccale, du pharynx et du larynx, à savoir la respiration, la déglutition et la phonation, auxquelles s'ajoute la fonction sphinctérienne du larynx.

Nous nous attacherons ici à la description de la prise en charge orthophonique sur les versants de la parole et de son intelligibilité. En conséquence, c'est le niveau supra-glottique qui sera visé, l'intégrité du larynx –dont les cordes vocales- ayant été préservée ; l'émission de la voix est assurée. Le timbre et l'intonation seront évidemment modifiés par rapport à leurs caractéristiques préopératoires en raison des changements de configuration d'une partie des résonateurs. La vitesse de coordination des mouvements articulatoires sera aussi affectée à des degrés divers par la chirurgie et par les difficultés de mobilisation des structures reconstruites ou restantes.

3.3 Procédés de réhabilitation et stratégies d'adaptation

Sans entrer dans les détails [se reporter aux références], citons quelques procédés spécifiques de notre réhabilitation: les mouvements minimaux [9], la proprioception, la "Méthode Feldenkrais" [10], et la programmation de l'action [11]. Les exercices s'effectuent sans douleur, avec une amplitude des mouvements minimale, yeux ouverts ou yeux fermés.

Nous rechercherons un rendement musculaire optimal au niveau des articulateurs avec l'alternance des séquences d'activité et de repos, l'endurance avec la réalisation d'un plus grand nombre d'exercices, l'augmentation de la résistance et de la puissance musculaire de façon très progressive, sans fatigue et toujours sans douleur.

Première étape : la respiration

La canule de trachéotomie, située au niveau des premiers anneaux de la trachée, est de type non fenêtrée et avec ballonnet au début, pour assurer la ventilation du patient sans risque de fausse route à la déglutition incontrôlée de la salive, sa "langue" ne pouvant bouger. Cette canule empêche le passage de l'air au-dessus d'elle, à travers le larynx donc entre les cordes vocales, rendant impossible la production de la voix. Notre travail se situe en amont de la reprise de la phonation rendue possible quelques jours plus tard avec la pose d'un système fenêtré: l'air expiré sort en partie par les voies aériennes supérieures ; le ballonnet préserve des fausses routes et assure la protection des voies aériennes inférieures.

En premier lieu, le patient doit pouvoir respirer librement. Puis c'est le contrôle du flux d'air nécessaire à la vocalisation, à l'émission des consonnes plosives ou des constrictives, et à la répartition des pauses dans le discours qui est visé. Par exemple, nous exerçons au début l'apnée après inspiration et plus tard nous utiliserons la pression et la poussée de l'air pour la résistance musculaire au niveau labial.

Deuxième étape : les praxies

L'entraînement des praxies bucco faciales, c'est-à-dire "d'un point de vue physiologique la coordination des mouvements dans un but donné" [12], pour nous celui de la parole, aura pour objectif la prononciation et la différenciation futures de voyelles et de consonnes. Ce sont d'abord de légers mouvements de rapprochement des lèvres en vue de la fermeture de la cavité buccale versus son ouverture dans le sens vertical, d'étirement latéral des commissures versus la position de repos dans le sens horizontal, et de protrusion des lèvres versus leur position de repos dans le sens antéropostérieur. Nous préparons la production des consonnes bilabiales [p, b, m], des labiodentales [f, v], et des voyelles.

Le flux d'air au cours de la parole est libre (tractus vocal ouvert), limité (tractus vocal rétréci) ou bloqué (occlusion du tractus vocal). Nous recherchons en conséquence le rétrécissement du conduit vocal pour un passage d'air limité (production des phonèmes constrictifs), par exemple avec le rapprochement des lèvres ou de la mandibule et du maxillaire. Pour favoriser la réalisation des sons apico-dentaux, nous travaillons le contact "langue"-palais par élévation de la mandibule.

Troisième étape : les clics

L'obtention de clics, dans son acception orthophonique, ou "sons produits dans la cavité buccale par un mouvement de succion ou d'expulsion délimité par deux points d'occlusion"[12] présente en phonétique clinique et dans notre cadre de parole pathologique un intérêt particulier. En effet, l'exécution d'un clic peut être directement reliée à une "pré"-articulation de consonnes. Ce bruit est généré par et dans une cavité buccale traumatisée, avant toute émission vocale possible (se reporter plus haut à l'étape de la respiration) grâce à l'air résiduel de la bouche. Les clics requièrent des occlusions en elles-mêmes qui ici seront d'abord bilabiales et ensuite effectives au niveau de la "langue" en contact avec le palais. Enfin, la réalisation d'une expulsion ou d'une succion suppose le contrôle de la pression de l'air, de l'action et de la force musculaires.

Quatrième étape : le voisement des consonnes

À l'occasion de deux consultations de suivi médical, nous insisterons sur la différenciation et sur l'émission des consonnes fricatives ou occlusives non voisées et voisées [s], [z], [p], [b], [t], et [d]. En effet, les consonnes sonores semblent assourdies dans le discours et également en répétition de séquences VC et VCV (V=a).

4. MÉTHODE ET PROCÉDÉS D'ÉVALUATION DE L'INTELLIGIBILITÉ DE LA PAROLE

4.1 Matériel et conditions d'enregistrement

Trois enregistrements de conversation spontanée d'une durée de 25 secondes chacun ont été réalisés, l'un avec un micro unidirectionnel Sennheiser E 845 S sur un enregistreur Marantz Professional Stereo PMD 60

(fréquence d'échantillonnage = 44,1 kHz) et les autres avec camescope numérique Panasonic NV GS 15 sur cassette mini D.V.. Le patient se trouve dans une pièce de consultation calme qui lui est familière, un mois et demi, deux puis cinq mois après l'intervention.

4.2 Protocole d'évaluation de la parole

Le protocole utilisé [13] permet une évaluation de la parole dans une situation de conversation spontanée, incluant une appréciation globale de l'intelligibilité, son débit, sa prosodie et son articulation selon une échelle cotée de 0 (normalité) à 4 (altération massive), selon la gradation suivante : 1, proche de la normale, 2, altérée, et 3, très altérée. L'évaluation est effectuée dans une salle calme, par un jury d'écoute composé de médecins O.R.L., d'internes, et d'infirmières, auxquels les paramètres de la parole à apprécier ont été expliqués. Il s'agit donc d'un jury non expert et en conséquence pas excessivement exigeant quant à la qualité des productions.

5. RÉSULTATS ET INTERPRÉTATION

5.1 Résultat essentiel de la réhabilitation

Dans ce cadre défini de soins et de contexte complexes, la précocité de la réhabilitation orthophonique, commencée au quatrième jour post-opératoire semble essentielle. En moyenne, elle débute à 5 semaines, et au plus tôt à deux semaines postopératoires, pour un nombre moyen de 10 séances effectives, sur un total de 16 proposées [8]. Dans notre cas, 11 ont été réalisées.

Le patient, dès le port d'une canule fenêtrée, au quinzième jour post-opératoire, a recouvré une parole de qualité suffisante pour s'exprimer sans nécessité de recours à l'écriture ou à une autre personne familière pour se faire comprendre par son entourage. Dans des cas similaires de glossectomie totale avec préservation du larynx, la parole reprend dans un délai plus long et comporte au début des mots isolés ou des phrases courtes [6].

5.2 Étude détaillée des résultats de la réhabilitation

Nous constatons l'obtention de la fermeture buccale et de façon concomitante celle d'un léger clic bilabial en succion au bout d'une semaine d'entraînement des praxies. Deux jours plus tard, l'expulsion de l'air résiduel buccal prépare l'explosion de la consonne occlusive sourde [p], et assure sa prononciation. D'autre part, la "langue" vient au contact des bords latéraux du palais. Dans les jours qui suivent, le clic bilabial devient plus audible : nous supposons que la force musculaire augmente. En même temps, un clic faible mais audible est produit en aspiration de l'air [tsst] à l'occlusion "langue"-palais. La production du voisement des consonnes sonores [b], [d] et [z] apparaît isolément ou dans une séquence VC ou VCV, (V=a), mais pas dans la parole.

5.3 Résultats de l'évaluation

Les résultats de l'évaluation de la parole sont mis en évidence et analysés avec le calcul des taux de distorsion.

L'échelle de cotation de l'intelligibilité et des paramètres de la parole comporte 5 degrés, notés de 0 à 4, ramenés pour le calcul de la moyenne de 1 à 5 puis convertis en pourcentages. Plus ceux-ci sont élevés, plus les altérations sont importantes. Globalement, les taux de distorsion (Fig.1) s'améliorent dans l'intervalle d'un mois et demi (M1.5), à 5 mois (M5) de l'intervention, après une dégradation entre M1 .5 à M2 (deux mois après chirurgie).

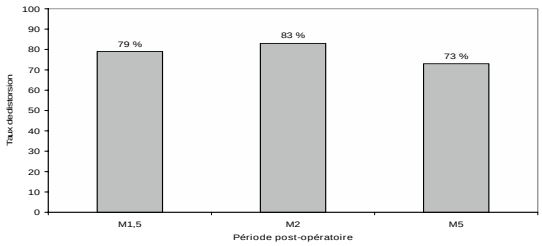


Figure 1 : Évolution du taux de distorsion de la parole.

L'observation détaillée des domaines explorés (Fig. 2) montre, de M1.5 à M5, une stabilité de l'altération des éléments suprasegmentaux de la parole, avec 60% de distorsion pour le débit, et 68% pour la mélodie. La qualité de l'intelligibilité et surtout celle de l'articulation évoluent nettement, en passant de la catégorie "altération massive" à celle de "très altérée" de M1.5 à M5, soient des taux respectivement de 92% et 84% pour l'intelligibilité, et de 96% à 80% pour l'articulation.

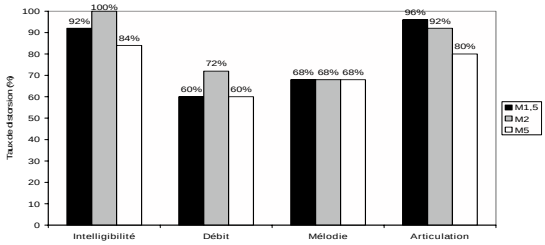


Figure 2 : Évolution des taux de distorsion de l'intelligibilité, du débit, de la mélodie et de l'articulation, à M1.5, M2 et M5 post-opératoires.

5.4 Interprétation et discussion

Malgré une chirurgie sévère et des particularités du tractus vocal caractérisé par un prognathisme mandibulaire, une proalvéolie et un palais ogival, nous avons proposé une réhabilitation précoce de la phonation, basée et organisée à partir des caractéristiques phonétiques de la parole. La mobilisation des structures anatomiques restantes ou reconstruites a été sollicitée en phase post-opératoire immédiate, sans préjuger des résultats. Les caractéristiques morphologiques de l'articulé temporo-mandibulaire n'ont pas empêché l'occlusion bilabiale, et celle de la "langue" avec le palais.

Dans les suites thérapeutiques, le patient a suivi une radiothérapie d'une durée de sept semaines, commencée la veille de notre premier enregistrement à M1.5. Les effets secondaires de la chirurgie et de l'irradiation se traduisent par une fibrose musculaire, rigidifiant les muscles, moins mobiles, donc gênant les mouvements des articulateurs et leur vitesse. L'étude de Furia et al. [8] mentionne

l'importance de la réhabilitation en cours de radiothérapie pour maximaliser la mobilité des articulateurs. En conséquence, nous pouvons supposer, sans prendre en compte les capacités de compensation spontanées du patient, que le débit, l'intelligibilité et l'articulation de la parole se dégradent au fur et à mesure des enregistrements. Effectivement, (voir Fig. 2), le taux de distorsion du débit passe de 60 à 72% de M1.5 à M2, pour revenir à 60% trois mois plus tard, soit une altération légèrement supérieure à une altération moyenne. Remarquons simplement que ce ne sont pas deux semaines d'irradiation qui peuvent induire une telle différence de souplesse musculaire à M2.

En ce qui concerne l'intelligibilité, une dégradation apparaît d'abord dans des proportions moins importantes que pour le débit : le taux de distorsion est majoré de 8% entre M1.5 et M2 (voir Fig. 2). La mélodie restant stable et l'articulation s'améliorant dans cet intervalle, nous pouvons émettre une hypothèse de corrélation de la baisse de la qualité de la prononciation avec celle de l'intelligibilité à deux niveaux : la dégradation du débit est moins ressentie dans l'intelligibilité, du fait de l'amélioration de l'articulation. Le lambeau de reconstruction myo-cutané comporte des tissus musculaires et adipeux. Ceux-ci fondent progressivement, sur une durée d'environ deux à trois mois. Le volume de la "langue" diminue, ce qui entrave l'occlusion "langue"-palais et l'ensemble des réalisations articulatoires autres que les consonnes bilabiales ou labio-dentales. Même si le patient n'a pu poursuivre les exercices de rééducation qu'il réalisait seul auparavant, il a pu trouver de lui-même des compensations articulatoires [14] qui n'ont pas pu être analysées.

Dans notre cas de travail pensé par rapport et en préalable à la production et à la perception de la parole, c'est bien l'évolution de la performance motrice qui atteste d'une diminution constante du pourcentage des déformations articulatoires, successivement de 96, 92 à 80%. Les résultats des études menées sur l'impact d'une rééducation orthophonique soulignent tous l'amélioration de la parole et de son intelligibilité [1, 8, 15].

Bien sûr, cette évaluation de la parole et de son intelligibilité, proposée ici dans un cas unique de chirurgie rare, par un jury non expert, demande d'être complétée de façon objective par une analyse de l'habileté articulatoire du patient.

6. CONCLUSION

La technique chirurgicale de reconstruction décrite dans un cas de cancer de la cavité endobuccale, grâce à la conservation de l'intégrité du larynx, a pu préserver la production de la voix. La parole, malgré la modification importante de la géométrie du conduit vocal, ne montre pas d'altération majeure de son débit et de sa mélodie. Notre étude, illustrée de façon précise par les enregistrements et leur évaluation, souligne l'intérêt d'une prise en charge orthophonique avancée en phase post-opératoire immédiate, ciblée d'emblée sur les mouvements articulatoires, avant la reprise de la phonation. La qualité de l'articulation et de son intelligibilité s'améliore après réhabilitation orthophonique précoce. Notre démarche, dans ce contexte de traitement invalidant des articulateurs ouvre une nouvelle perspective thérapeutique.

Remerciements

Nous remercions particulièrement les membres du service O.R.L. de l'hôpital de Hautepierre pour leur collaboration. Ce travail est financé en partie par un Programme ACI TTT du Ministère de la Recherche, 2003-2006.

BIBLIOGRAPHIE

[1] R.C. Donaldson, M. Skelly, F.X. Paletta. Total glossectomy for cancer. *American Journal of Surgery*, 116:585-590, 1968.

[2] P.M. Kothary, J.C. Paymaster, G.G. Potdar. Radical total glossectomy. *British Journal of Surgery*, 61:209-590, 1974.

[3] M.Z. Effron, J.T. Johnson, E.N. Myers et al. Advanced carcinoma of the tongue. *Archives of Otolaryngology*, 107:694-697, 1989.

[4] M.R. Sultan, J.J. Coleman. Oncologic and functional considerations of total glossectomy. *American Journal of Surgery*, 158:297-302, 1989.

[5] R.S. Weber, L. Ohlms, J. Bowman et al. Functional results after total or near total glossectomy with larungeal preservation. *Archives of Otolaryngology Head and Neck Surgery*, 117:512-515, 1991.

[6] R. Tiwari, A. Karim, A.-J. Greven and B. Gordon. Total glossectomy with laryngeal preservation. *Archives of Otolaryngology, Head and Neck Surgery*, 119: 945-949, 1993.

[7] C.M. Ruhl, L.L. Gleich, J.L. Gluckman. Survival, Function and Quality of Life After Total Glossectomy. *The Laryngoscope*, 107: 1316-1321, 1997.

[8] C. Furia, L. Kowalski, M. Latorre, E. Angelis, N. Martins,. A. Barros and K. Ribeiro. Speech intelligibility after glossectomy and speech rehabilitation *Archives of Otolaryngology, Head and Neck Surgery*, 127: 877- 883, 2001.

[9] W. Weiss. *La voix mobile.* Masson, Paris, France, 1994.

[10] M. Feldenkrais. *Énergie et bien-être par le mouvement*. Dangles, St Jean de Braye, France, 1993.

[11] S. Dehaenne. *Le cerveau en action.* Presses Universitaires de France, Paris, France, 1997.

[12] © Encyclopædia Universalis 2005, tous droits réservés

[13] M. Calmet-Smadja. Évaluation de la qualité de vie après glossectomie partielle : étude des corrélations entre la qualité de vie, l'évaluation fonctionnelle et l'intelligibilité de la parole. *Mémoire d'orthophonie*. Université Pierre et Marie Curie, Paris VI, 2003.

[14] C. Savariaux, P.Perrier, J. Lebeau, G. Magaña, C. Dorange-Pattoret. Production de parole après traitements de cancers de la cavité buccale. In *Actes des XXIIIèmes Journées d'Étude sur la Parole*, pages 433 - 436, 2000.

[15] J. Teichgraeber, J. Bowman, H. Goefert. New test series for the functional evaluation of oral cavity cancer. *Head and Neck Surgery*, 8: 9- 20, 1985.

Evolution de la perception des phonèmes, mots et phrases chez l'enfant avec Implant Cochléaire : un suivi de trois ans post-implant

Victoria MEDINA [1,2,3] & *Willy SERNICLAES* [2]

1 UFR- Linguistique, Université Paris 7, Denis Diderot, 2 Place Jussieu – 75251 Paris Cedex 5
2 Lab. Psychologie de la Perception, CNRS, Université René Descartes, Paris 5
3 CTNERHI, 75013 Paris
E-mail : medina_vicky@yahoo.fr

ABSTRACT

The aim of the present study was to examine the development of the perception of phonemes, words and sentences in a group of 18 children with cochlear implant (IC) from 12 to 36 months after implantation. The results show that the perceptual development of the different segments is fairly linear, that the rate of development is faster for words vs. phonemes and for sentences vs. words and that consonant perception, but not vowel perception, predicts later development of word and sentence perception.

1. INTRODUCTION

L'implant cochléaire (IC) est une prothèse électro-acoustique ayant pour objectif pédiatrique de permettre à l'enfant sourd profond accéder à la communication linguistique.
L'apport de l'IC pour la réhabilitation de la surdité profonde pré -linguistique est manifeste [1], mais il est également très variable. Parmi quelques facteurs responsables de la variabilité interindividuelle on relève l'âge au début de la surdité, la durée de surdité sans implant, la durée d'utilisation de l'implant, le type de processeur, le mode de communication [2] et l'âge d'implantation [3]. L'effet de ces différences sur le développement de la production et de la perception de la parole s'explique par leur incidence sur le niveau de structuration phonologique avant implantation [4].

La complexité linguistique dans les processus de communication parlée est mesurable à partir de la quantité d'information véhiculée par le signal [5]. Le locuteur ajuste tant la complexité linguistique que la qualité acoustique du signal pour que celui-ci soit correctement décodé. La théorie H&H (Hypo et Hyper Speech) [6] dit que le locuteur assure la quantité d'information transmise en fonction de la loi du moindre effort. En d'autres termes, le locuteur cherche à être compris, mais pas à n'importe quel prix. Si le message à transmettre est particulièrement simple, prédictible par l'auditeur, de l'Hypo-Speech suffit au locuteur pour le faire passer. Nous savons par exemple que la contribution du contexte linguistique à l'intelligibilité des mots dépend de la dégradation acoustique [5, 7].
L'information sensorielle issue du signal acoustique active successivement différents niveaux de traitement (phonologique, lexical, syntaxique, sémantique) avec des rétroactions haut-bas ('top-down') dont la nature et l'étendue diffère selon le modèle envisagé (TRACE [8] ou MERGE [9]).

Dans le cas des enfants avec Implant Cochléaire (IC) l'intelligibilité du signal de parole s'améliore dans la mesure où l'enfant s'adapte à l'IC. La contribution des différents niveaux de traitement au progrès des performances de communication parlée avec IC n'a cependant pas été abordée jusqu'à présent.

L'objectif de cette étude est d'examiner l'évolution de la perception des phonèmes, mots et phrases entre 12 et 36 mois chez des enfants sourds appareillés avec IC. Les questions suivantes ont fait l'objet d'une attention particulière : (1) Le développement de la perception de ces différents segments de parole est-il linéaire? (2) Le rythme de développement est-il indépendant de la complexité du segment ? (3) Le développement des segments de niveau supérieur (phrases ou mots) est-il prédictible à partir de celui des segments de niveau inférieur (mots ou phonèmes) ?

2. MÉTHODE

2.1. Sujets

Un groupe de dix huit enfants sourds congénitaux avec IC (5 garçons et 13 filles) ont été testés à 12, 24 et 36 mois post-IC. Leur âge chronologique s'étalait entre 3 ans et 7 ans (D.S. : 1,5) avec une moyenne de 4 ans 8 mois (D.S. : 1,4) d'implantation à 12 mois post-IC. Six enfants ont été implanté avant 3 ans (âge moyen : 2 ans 4 mois) ; six enfants ont été implantés entre 3 et 4 ans (âge moyen : 3 ans 4 mois) et six enfants ont été implantés après 4 ans (âge moyen : 5 ans 6 mois).

Ces enfants font partie d'un suivi longitudinal [10] et ils ont été suivis dans 4 CHU (Lyon- Eduard Herriot, Montpellier- Saint Charles, Paris- Trousseau, Toulouse- Purpan).

Nous n'avons pas utilisé de critères d'exclusion en relation avec le type de l'implant cochléaire.

2.2. Procédure

Le protocole d'évaluation a été adapté du Test d'Evaluation des Perceptions et des Productions de la Parole (TEPP) [11]. Ce test permet une évaluation longitudinale des compétences auditivo-perceptives des enfants sourds de 2 à 10 ans.

« Identification des phonèmes. Voyelles et Consonnes »
Cette épreuve comportait deux séries, une vocalique de 16 voyelles isolées et une consonantique de 17 consonnes en contexte /a/. La tâche consistait à répéter la syllabe.

16 voyelles et glides :
[a], [i], [u], [o], [ɔ], [y], [e], [E], [ø], [œ], [ã], [õ], [Ẽ], [w], [j], [ɥ].

17 consonnes en contexte /a/ :
[p], [b], [m], [t], [d], [n], [k], [g], [ɲ], [f], [v], [s], [z], [ʃ], [ʒ], [l], [r].

« Identification des mots »
Identification de 12 syntagmes nominaux constitués d'un nom et d'un déterminant, avec un niveau de vocabulaire très simple et connu de l'enfant testé. La tâche consistait à désigner le mot adéquat sur une planche constituée d'images illustrant 12 mots répartis en : syntagmes disyllabiques, trisyllabique et polysyllabiques.

« Identification de phases simples »
Identification des phrases constituées de deux syntagmes, un nominal et un verbal, faciles à comprendre pour l'enfant. Chaque item présentait trois confusions possibles : sur le syntagme nominal, sur le syntagme verbal ou sur l'ensemble du message. Une série de 10 planches ont été proposées, chaque planche était constituée de 4 images. Sur chaque planche l'enfant avait le choix entre 4 propositions du type : « la fille rit, le garçon rit, le garçon dort, le chien sort ». La tâche consistait à désigner la phrase adéquate sur une planche.

« Identification de phases complexes »
Identification des phrases, constituées de trois syntagmes, deux nominaux et un verbal. L'enfant devait mémoriser l'ensemble de la phrase et différencier les mots phonétiquement proches, (ex. le garçon range la balle, le garçon range le bol, le garçon cache la balle, le garçon cache le bol...).
Une série de 8 planches ont été proposées, chaque planche était constituée de 4 images illustrant des phrases construites selon le modèle complexe : syntagme nominal sujet- syntagme verbal-syntagme nominal objet. La tâche consistait à désigner la phrase adéquate sur une planche.

3. RÉSULTATS

3.1. Linéarité du développement

Afin d'examiner la linéarité du développement des performances, les scores moyens -pour l'ensemble de 18 participants- d'identification correcte (p) on été transformés en Logit (p)= Log_n (p/(1-p)), ceci pour redresser effets « plancher » (vers 0%) et « plafond » (vers 100%). Les transformées Logit des scores de perception de phonèmes, de mots et de phrases sont présentés dans la Figure 1. Les scores ont ètè traités sur SPSS à l'aide d'une Régression Logistique avec l'âge IC comme variable indépendante et le score de réponses correctes pour les différents segments (5 niveaux) comme variable dépendante.

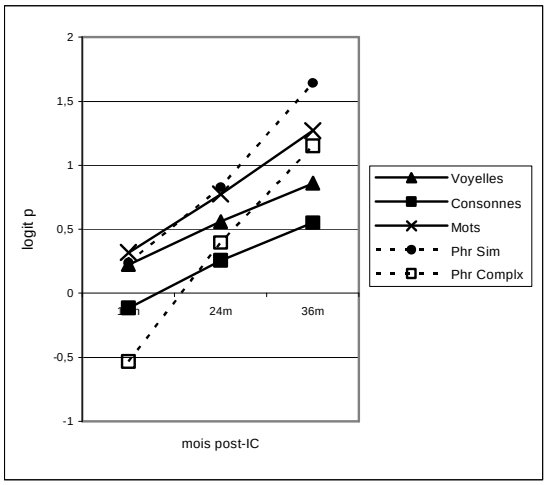


Figure 1 : Les transformées Logit des scores de perception de phonèmes, de mots et de phrases simples ou complexes à 12, 24 et 36 mois post-implant.

Les résultats montrent une évolution linéaire de la perception des voyelles, des consonnes, des mots et des phrases simples ou complexes. Le test de linéarité, obtenus en testant l'adéquation des fonctions de réponses correctes au modèle Logistique (test de Hosmer-Lemeshow) est non significatif pour chacun des 5 segments (p=.79, .56, .87, .46, et .18, respectivement pour les voyelles, consonnes, mots

phrases simples et complexes). L'effet de l'âge IC, tous segments confondus, est hautement significatif (Ξ^2 de Wald (dl=1)= 304, p<.001).

3.2. Parallélisme du développement pour les différents segments

La Figure 1 indique que les fonctions Logit correspondant aux voyelles et consonnes sont pratiquement parallèles. Par contre, la pente de la fonction est plus raide pour les mots et encore plus raide pour les phrases. L'interaction âge IC X segment est significative (Ξ^2 de Wald (dl=4)= 35, p<.001).
L'examen des contrastes d'interaction montre que la différence de pente entre voyelles et consonnes ainsi que celle entre phrases simples et complexes ne sont pas significatives (Ξ^2<1).
Par contre, la différence de pente entre voyelles et consonnes, prises conjointement, et mots ainsi que celle entre mots et phrases, simples et complexe prises conjointement, sont significatives (respectivement, Ξ^2 de Wald (dl=4)= 3.72, p=.05 ; 6.22, p<.05).

3.3. Prédictibilité des performances

Nous avons effectué des analyses de corrélation entre les différents segments (phonèmes, mots et phrases) prenant en compte les différents moments de passation des tests (soit 12, 24 ou 36 mois post-IC).

Pour les phonèmes (voyelles et consonnes, figure 1), il existe un effet de corrélation positive entre les voyelles à différents moments d'évaluation, par exemple, les voyelles à 12 mois et celles à 24 mois, (table 1). Il existe également un effet de corrélation positive entre les voyelles et les consonnes (p. ex. les voyelles à 12 mois et les consonnes à 12, 24 et 36 mois, voir table 1).

Pour les autres segments, les résultats montrent que l'intelligibilité des voyelles est corrélée avec celle des mots et des phrases sur la même période de passation. Les voyelles à 12 mois sont positivement corrélées avec les mots à 12 mois, les phrases simples à 12 mois et les phrases complexes à 12 mois post-IC (table 1). Par contre, les scores des voyelles à 12 mois ne sont pas corrélés avec ceux des mots ou phrases à 24 ou 36 mois. (p=.10 et .23, pour les mots à 24 et 36 mois respectivement; p=.23 et .25, pour les phrases simples à 24 et 36 mois respectivement; p=. 31 et .30, pour les phrases complexes à 24 et 36 mois respectivement).

Table 1: Corrélations entre Voyelles à 12 mois post-IC et différents segments.

	R^2	p
Voyelles 12M × Voyelles 24M	0.412	.004
Voyelles 12M × Consonnes 12M	0.460	.002
Voyelles 12M × Consonnes 24M	0.406	.004
Voyelles 12M × Consonnes 36M	0.331	.013
Voyelles 12M × Mots 12M	0.329	.013
Voyelles 12M × Phr. Simpl. 12M	0.263	.030
Voyelles 12M × Phr. Compl. 12M	0.384	.006

Les consonnes par contre sont corrélées avec les mots et les phrases sur les différentes périodes de passation. Nous observons une corrélation positive entre les consonnes à 12 mois et les mots à 12 mois, à 24 mois et à 36 mois (voir table 2) ; entre les consonnes à 12 mois et les phrases simples à 12 mois, à 24 mois et à 36 mois (table 2); et entre les consonnes à 12 mois et les phrases complexes à 12 mois, à 24 mois et à 36 mois post-IC (table 2).

Table 2: Corrélations entre Consonnes à 12 mois post-IC et différents segments.

	R^2	p
Consonnes 12M × Mots 12M	0.300	.019
Consonnes 12M × Mots 24M	0.230	.044
Consonnes 12M × Mots 36M	0.280	.024
Consonnes 12M × Phr. Simp 12M	0.227	.046
Consonnes 12M × Phr. Simp 24M	0.291	.026
Consonnes 12M × Phr. Simp 36M	0.309	.017
Consonnes 12M × Phr. Cmpl 12M	0.377	.007
Consonnes 12M × Phr. Cmpl 24M	0.315	.019
Consonnes 12M × Phr. Cmpl 36M	0.456	.003

Les voyelles et les consonnes à 36 mois ne présentent pas de corrélation avec d'autres segments dans la même période de passation.

La perception des mots garde un lien avec la perception des phrases. Les mots à 12 mois présentent une corrélation positive avec les phrases simples à 12 mois, à 24 mois et à 36 mois (voir table 3). Les mots à 24 mois ont un effet de corrélation avec les phrases complexes à 24 mois et à 36 mois (table 3).

Table 3: Corrélations entre Mots à 12 et 24 mois post-IC et Phrases.

	R^2	p
Mots 12M × Phr. Simp 12M	0.254	.033
Mots 12M × Phr. Simp 24M	0.234	.049
Mots 12M × Phr. Simp 36M	0.401	.005
Mots 24M × Phr. Cmpl 24M	0.339	.014
Mots 24M × Phr. Cmpl 36M	0.471	.002

L'évolution des phonèmes, des mots et des phrases est indépendante de l'âge d'implantation, la corrélation est non significative (p=.55, .74, .70, .06 pour les voyelles, consonnes, mots et phrases respectivement).

4. DISCUSSION ET CONCLUSIONS

Nos résultats montrent que l'amélioration de la perception se fait de façon linéaire pour les phonèmes, mots et phrases. Ceci met en évidence la continuité du développement perceptif sur cette période lorsque l'on tient compte des effets de seuils inhérents à la progression de toute variable catégorielle.
Cependant, la progression est plus rapide pour les phrases par rapport aux mots et pour ces derniers par rapport aux phonèmes. Ceci met en évidence une vitesse de développement plus rapide pour les segments les plus complexes et suggère un effet démultiplicateur de la perception des phonèmes sur celle des mots et des phrases.

Enfin, les résultats de cette étude pilote mettent en évidence des relation entre différents segments à différentes périodes d'évaluation de la perception post-IC. Il semblerait que la perception des voyelles à 12 mois post-IC permettrait de prédire la perception des mots et des phrases sur la même année d'évaluation post-IC. Par contre, la perception des consonnes à 12 mois post-IC permettrait de prédire la perception des mots et des phrases sur les différentes années d'évaluation (12, 24 et 36 mois post-IC). Ceci suggère que l'intelligibilité des mots et des phrases garde une relation plus forte au long des années avec celle des consonnes qu'avec celle des voyelles. Il semblerait qu'une bonne perception des consonnes à 12 mois post-IC permettrait de prédire une bonne perception des mots et des phrases jusqu'à 36 mois post-IC.

En conclusion, ces résultats mettent en évidence le caractère progressif du développement de la perception de différents segments de parole avec IC et suggèrent que l'intelligibilité des phonèmes, et surtout des consonnes, affecte celle des mots et des phrases avec des effets démultiplicateurs.

BIBLIOGRAPHIE

[1] C. Allen, T. P. Nikolopulos & G. M. O'Donghue. Speech intelligibility in children after cochlear implantation. *The American Journal of Otology, 19,* 742-746. 1998.

[2] R.T. Miyamoto, M.J. Osberger, S.L. Todd, A.M. Robbins, B.S. Stroer, S. Zimmerman-Phillips, A.E. Carney. (1995). Variables affecting implant performance in children. *Laryngoscope, 104*, 1120-1124. 1995.

[3] M. Svirsky, S-W. Teoh & H. Neuburger. (2004). Development of language and speech perception in congenitally, profoundly deaf children as a function of age at cochlear implantation. *Audiology & Neuro-Otology*, *9*, 224-233. 2004.

[4] W. Serniclaes, V. Medina, F. Schepers & P. Simon. Chapitre II. Le développement de la communication parlée avec Implant Cochléaire. *Surdité et langage : Prothèses, LPC et Implant Cochléaire.* En publication.

[5] C. Benoît & C. Abry. De l'impertinence, ou comment relier complexité linguistique et qualité acoustique. *Rapport de recherche de l'ICP,* 5, 179-189. 1995.

[6] B. Lindblom. Adaptative variability and absolute constancy in speech signals : Two themes in the quest of phonetic invariance. *Proceedings of the XIh International Congress of Phonetic Sciences, 3*, 9-18, Talinn, Estonia. 1987

[7] G. A. Miller & S. Isard. Some perceptual consequences of linguistic rules. *Journal Verbal Learning and Verbal Behaviour, 2*, 217-228. 1963.

[8] J. L. Mc Clelland & J. L. Elman. The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86. 1986.

[9] D. Norris, J.M. McQueen, & A. Cutler. Merging information in speech perception: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299-370. 2000.

[10] J. Sanchez, A. Bounot, V. Ansel. Suivi longitudinal sur dix ans d'enfants sourds pré-linguaux implantés, *Handicap - revue de sciences humaines et sociales*, n° 98, 63-70. 2003.

[11] A. Vieu, M. Mondain, M, Sillon, J.P. Piron et A. Uziel. Test d'Evaluation des Perceptions et Productions de la Parole (TEPPP). *Revue de Laryngologie, Otologie et Rhinologie*, *120*, 219-225, 1999.

Etude de la dysprosodie parkinsonienne: Analyses acoustiques d'un schéma de type interrogatif

Karine Rigaldie, Jean Luc Nespoulous, Nadine Vigouroux

Laboratoire Jacques Lordat, Toulouse
Institut des Sciences du Cerveau, Toulouse
IRIT, UMR, CNRS, Université Paul Sabatier, Toulouse
{rigaldie; vigourou}@irit.fr; nespoulo@univ-tlse2.fr

ABSTRACT

This article aims to acquire a better knowledge of prosodic disturbances in Parkinson disease via an acoustic analysis. The investigation of the patients' vocal productions by the way of acoustic analyses should indeed allow to identify phonetic and prosodic parameters that are specific of such a pathology. The Parkinsonian subjects had to repeat the interrogative pattern "Vous avez appris la nouvelle?" (in english: "You heard the news?"), three times: at the beginning, in the middle, and in the end of the protocol. This timing was determined in order to evaluate the effects of tiredness and the influence of other stimuli during the protocol. In order to determine the effect of dopamine, oral productions of twelve parkinsonian patients have been collected, in the OFF and ON states, and have then been compared to those of control subjects.

1. INTRODUCTION

Selon Chevrié Muller [1], les altérations prosodiques des sujets parkinsoniens seraient dues à des troubles de la réalisation motrice de la parole. Et pour Darley [2], l'origine de la dysarthrie parkinsonienne s'expliquerait par la limitation de l'exécution des mouvements respiratoires et phonatoires, liée à une faiblesse de la rigidité musculaire.

Au niveau de la parole dite « normale », la prosodie se traduit par des variations d'indices acoustiques et permet une organisation linguistique de l'énoncé. La dysprosodie est donc une manifestation de la dysarthrie parkinsonienne. Les manifestations de ces troubles prosodiques sont : une articulation appauvrie manifestée par une faible intensité (palilalie), un débit de parole accéléré (tachylalie) et un bredouillement lors du démarrage de la production ainsi que des variations de débit (dysfluence). La dysprosodie est souvent le premier signe de la dysarthrie parkinsonienne [2] et semble résister aux traitements médicamenteux [3].

Ce n'est que dans les récents travaux de Benoît Lagrue et Bernard Teston [4] et [5] que ces troubles de la parole ont été étudiés au moyen d'analyse acoustique. Excepté ces travaux, et ceux de Gentil [6], l'état de l'art montre que peu d'études sur la dysprosodie parkinsonienne ont été conduites en français; dans les deux conditions état OFF (sans médicament) et état ON (de prise médicamenteuse).

L'un des objectifs relatif à l'étude d'un schéma intonatif de type interrogatif répliqué trois fois, au début, au milieu et en fin de protocole, est notamment d'observer les variations de la fréquence fondamentale et de l'énergie, ceci en état ON et OFF de la prise médicamenteuse des sujets.

En 1998, Le Dorze [7] observent les variations du F0 sur des schémas intonatifs de type question-affirmation. Ils comparent le F0 de la dernière syllabe sur chaque phrase et observent que 10 patients produisent un F0 moins élevé comparé aux sujets de contrôle. Tandis que les sujets normaux réalisent les productions avec un F0 plus élevé sur la dernière syllabe, les sujets parkinsoniens ne montrent aucune variation.

2. BASE D'INFORMATIONS MÉDICALES ET DE DONNÉES DE PAROLE

Nous disposons d'une base de données médicale, neurologique et de parole constituée en collaboration avec l'équipe du CIC (Centre d'investigation Clinique) de l'hôpital de Purpan dans le cadre d'un projet soutenu par l'INSERM. La base de données comporte plusieurs types d'énoncés selon le protocole de Laur et Vigouroux [8].

2.1. Les Patients

La présente étude porte sur un corpus de parole recueilli auprès de 12 patients parkinsoniens en état OFF et ON (cinq femmes : **P1, P2, P3, P4** et **P5** et sept hommes : **P6, P7, P8, P9, P10, P11et P12**) aux degrés 2.5, 3 et 4 de l'échelle de Hoehn et de Yahr [8] et de 12 sujets de contrôle (sept femmes : **S1, S2, S3, S4, S5, S6** et **S7** et cinq hommes : **S8, S9, S10, S11** et **S12**). Afin de rendre compte de la variabilité inter-sujets nous avons comparé les productions des sujets parkinsoniens en état OFF et ON de prise médicamenteuse et les avons comparées à celles des sujets de contrôle.

Les patients parkinsoniens retenus sont d'origine française, ont entre 60 et 80 ans et présentent au plan de la perception une altération de la parole repérée par l'équipe médicale selon l'échelle de la maladie de Hoehn et de Yahr [9].

2.2 Le protocole

Les patients sont convoqués la veille de l'investigation. Ils interrompent leur traitement 16 heures avant la première passation et sont enregistrés le lendemain matin vers 8H30. A la fin de la première série d'exercices (phonétiques), ils prennent leur traitement habituel et peuvent se reposer. Dès qu'ils sont « débloqués », c'est-à-dire dès qu'ils ont atteint la phase ON, ils répètent exactement la même procédure.

3. METHODOLOGIE

Le schéma interrogatif correspond à une question de type syntaxique déclaratif, "Vous avez appris la nouvelle?". La phrase interrogative devrait être marquée par une montée importante de la voix à la fin de la phrase surtout quand la structure de la phrase est de type énonciatif, comme c'est le cas dans notre étude. La forme de la pente mélodique est déterminante. Plus l'angle se rapproche de 90°, plus la courbe est perçue comme une question, Léon [10]. Il faut souligner que les formes des courbes peuvent présenter de nombreuses variantes selon les caractéristiques intra et inter individuelles du locuteur et des effets phonostylistiques comme par exemple un niveau d'attaque plus ou moins haut pour une question, une montée plus ou moins ample pour une continuation, une descente plus ou moins brusque pour une finalité.

Le stimulus « vous avez appris la nouvelle ? » a donc été produit trois fois par l'ensemble des sujets femmes et hommes, ceci en début du protocole N1 (première production de nouvelle), au milieu du protocole : N2 et à la fin du protocole : N3. Nous avons posé l'hypothèse selon laquelle les valeurs de fréquence, d'intensité et de durée de la dernière occurrence N3 seraient plus basses comparées à celles de la première N1 et de la deuxième N2. Cela traduirait selon nous un effet de fatigue. Pour l'étude des différents paramètres acoustiques, et afin d'observer la meilleure production entre les deux états des patients, nous avons procédé à deux analyses, nous avons tout d'abord comparé la moyenne des trois occurrences chez tous les sujets puis avons comparé chacune de ces trois occurrences entre elles.

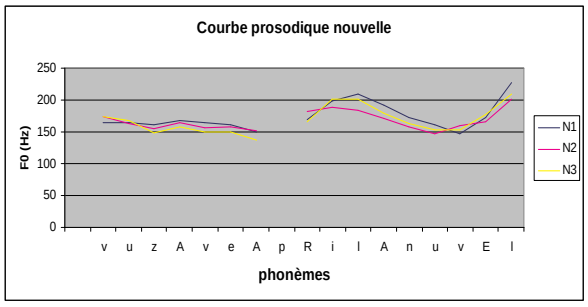


Figure 1: Représentation prosodique de N1, N2, N3, Etat OFF pour P6

A partir des marques d'annotation, les valeurs moyennes de durée, de fréquence fondamentale, des formants des phonèmes ont été calculées. Pour cela, nous avons également utilisé le logiciel Winsnoori (Loria Babel Technologie, 2002). L'algorithme de calcul du F0 est basé sur la méthode : Fast Fourier Transform (FFT). Le signal est digitalisé à 22 KHz. Pour le F0 et l'énergie, nous avons utilisé un pas de discrétisation de cinq millisecondes. Les valeurs de F0 sont calculées toutes les cinq millisecondes (en utilisant l'algorithme).

La gamme de variations moyennes de la fréquence fondamentale se situe aux alentours de 100 à 150 Hz pour l'homme adulte et de 140 à 240 Hz pour la femme adulte. Cependant au niveau de la parole pathologique, ces moyennes peuvent varier considérablement. Nous avons donc comparé les résultats de la population des sujets femmes et celle des sujets hommes séparément afin de pouvoir les interpréter au mieux.

Nous présenterons donc les moyennes du F0, de l'intensité pour chacune des populations, puis détaillerons les résultats obtenus chez les sujets en état OFF et ON ainsi que chez les sujets sains.

4. RESULTATS

Pour une meilleure lisibilité des graphiques lors de la présentation des différents résultats et notamment des moyennes, nous avons regroupé les résultats des patientes parkinsoniennes en état (PF OFF), des patientes parkinsoniennes en état ON (PF ON) et des sujets de contrôle femmes sous le terme générique de « sujets femmes ». Le terme de « sujets hommes » prend en compte la comparaison des résultats obtenus entre les productions des patients parkinsoniens en état (PH OFF), des patients parkinsoniens en état ON (PH ON) et des sujets de contrôle hommes.

Afin de mesurer les variations de la fréquence fondamentale, de l'intensité et de la durée entre l'état OFF et l'état ON, nous avons en premier lieu moyenné les trois productions du schéma de type interrogatif chez l'ensemble des sujets. Nous constatons ainsi que le traitement pharmacologique améliore l'ensemble des productions des patients **P2**, **P4** et **P5**. Un seul patient

n'améliore pas l'ensemble des ces valeurs entre l'état OFF et l'état ON, il s'agit du patient **P8.**

4.1. Sujets femmes

Variabilité intra-groupes : le cas du F0

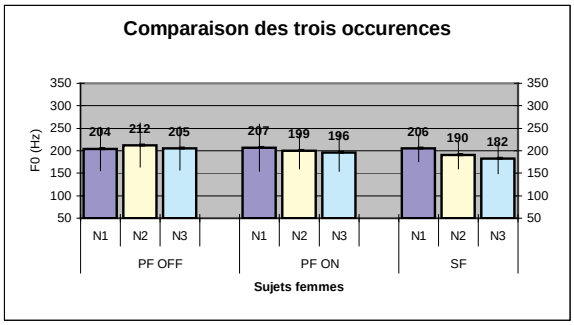


Figure 2 : F0 sujets femmes

Nous avons étudié les valeurs de N1, N2 et N3 sur l'ensemble des sujets femmes. Il ressort que N3 est la plus mauvaise réalisation sur l'ensemble des patientes en état ON (206,79, 199,35 et 196,19 Hz) et chez les sujets de contrôle femmes (205,50 Hz ; 190,01 Hz et 182,34 Hz). Chez les patientes en état ON, N1 est la meilleure réalisation.

En ce qui concerne le paramètre de F0, l'effet de fatigue semble donc se vérifier chez trois patientes en état ON (**P1**, **P2** et **P4**), trois sujets de contrôle (**S1**, **S5** et **S7**) et une patiente en état OFF (P4) puisque les valeurs déclinent entre la première, la deuxième et la troisième réalisation. Nous remarquons une augmentation entre N1 et N2 chez trois patientes en état OFF mais les valeurs diminuent entre N2 et N3, il s'agit de (**P1**, **P2** et **P4**). Ainsi quel que soit leur état la dernière réalisation est toujours la plus mauvaise chez ces patientes. Mais la dopamine améliore l'ensemble des trois réalisations chez **P2** et **P4**.

Variabilité intra-groupes : le cas de l'énergie

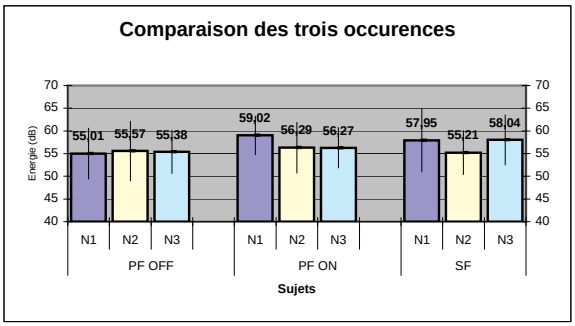


Figure 3 : Energie sujets Femmes

Le paramètre de l'intensité entre les trois occurrences semble plus homogène en état OFF. Avant la prise de dopamine la moyenne de l'énergie de la première occurrence est de 55,01dB, 55,57 dB pour la deuxième et 55,38 dB pour la dernière. En état ON la moyenne est plus basse sur la dernière réalisation, 59,02 dB pour N1, 56,29 dB pour N2 et 56,27 dB pour N3. Chez les sujets femmes, c'est la deuxième réalisation qui est produite avec une plus faible intensité 55,21 dB pour N2, 57,95 dB pour N1 et 58,04 dB pour N3.

La patiente **P4** est la seule patiente dont les valeurs de fréquence diminuent entre chaque occurrence alors que ces valeurs d'intensité augmentent entre ces mêmes occurrences et ce quel que soit son état. Cette patiente semble donc compenser son déficit au niveau des vibrations des cordes vocales par une élévation de l'énergie. Chez les autres patientes les valeurs d'intensité et de fréquence diminuent entre chaque occurrence en état ON. L'effet de compensation du déficit du F0 se vérifie également chez deux sujets de contrôle femme (**S1** et **S7)**.

4.2. Sujets hommes

Variabilité intra-groupes : le cas du F0

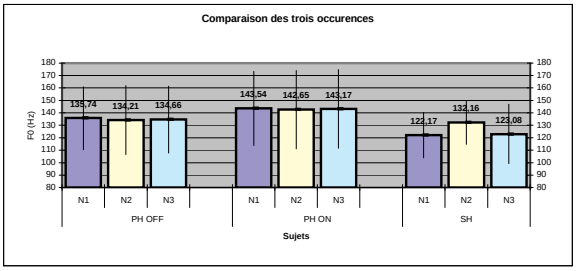


Figure 4 : F0 sujets Hommes

La figure 3 indique que sur l'ensemble de la population patients parkinsoniens état OFF, état ON et sujets de contrôle, la moyenne des trois occurrences est plus homogène chez les patients en état ON. Que ce soit en OFF ou en état ON la deuxième réalisation est toujours la plus mauvaise chez les patients parkinsoniens (135, 74Hz pour N1, 134, 21 Hz pour N2 et 134, 66Hz pour N3 en OFF et 143, 54 Hz pour N1, 142, 65 Hz pour N2 et 143, 17 Hz pour N3 en état ON).

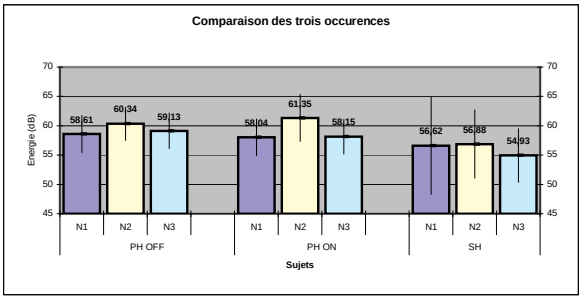


Figure 5 : Energie Sujets Hommes

Variabilité intra-groupes : le cas de l'énergie

Les valeurs d'énergie sont plus élevés sur la deuxième occurrence que se soit en état OFF ou ON. Nous avions constaté dans le paragraphe précédent que N2 était la

production la moins bien réalisé au niveau des variations du F0 sur l'ensemble des patients et ce quel que soit leur état. Cela suppose que les sujets parkinsoniens hommes compenseraient leur déficit du F0 par une élévation de l'énergie sur la deuxième occurrence.

5. CONCLUSION

Les résultats obtenus sur le stimulus « vous avez appris la nouvelle ? » nous permettent de vérifier plus précisément les deux principales hypothèses relatives à ce stimulus à savoir le respect ou non du schéma mélodique de base (réalisation ascendante en finale de production et ce sur l'ensemble des trois occurrences) par l'ensemble des patients et des sujets et l'observation ou non d'un effet de fatigue sur la troisième occurrence. Nous avons ainsi noté que l'ensemble de la population (hommes et femmes) ne présentaient pas de difficulté à réaliser un schéma de type interrogatif.

Nous avions également posé l'hypothèse que la troisième réalisation serait la plus mauvaise tant au niveau des valeurs du F0 que celles de l'énergie de l'énergie, traduisant un certain effet de fatigue.

En ce qui concerne le paramètre de fréquence fondamentale et en état OFF nous observons cet effet de fatigue chez les patients **P4**, **P8** et **P11** entre la première, la deuxième et la troisième réalisation. En état ON, ce phénomène apparaît chez **P8.**

Les valeurs d'énergie quant à elles diminuent entre N1, N2 et N3 chez **P2**, **P5** en état OFF et chez **P1**, **P2** et **P3** en état ON. Nous avons également remarqué sur ce stimulus, un effet de compensation du déficit des vibrations des cordes vocales par une élévation des valeurs de l'énergie. En effet, alors que les valeurs de fréquence déclinent entre chaque occurrence les valeurs d'énergie augmentent. Nous avons observé ce phénomène chez les patients **P4** et **P11** en état OFF **et** chez **P7, P8 et P9** en état ON**.** Ceci s'observe également chez les sujets de contrôle **S1**, **S7**, **S8** et **S9**.

Un troisième niveau d'étude concerne le paramètre de durée. Nous avons initié une analyse de la durée de la production et du débit.

En effet, nous formulons également l'hypothèse selon laquelle la dopamine pourrait avoir, un effet sur le débit ou la vitesse d'articulation des productions des sujets. De plus, nous pensons que l'augmentation ou le ralentissement du débit serait donc tout comme l'intensité une stratégie consciente ou non mise en œuvre par le patient pour compenser le déficit en termes de vibration des cordes vocales.

BIBLIOGRAPHIE

[1] Chevrie Muller C., Intervention rééducative sur la voix et la parole dans la maladie de Parkinson, in *La maladie de Parkinson*, Acanthe (ed)) Masson, Paris, A. Rascol, pp. 223-237, 1998.

[2] Darley, F.-L., Aronson, A.-E., Brown, J.-R. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, **12**, pp. 249-269, 1969.

[3] Shea, B., Drummond, S., Metzer, W., et Kreuger, K. Effect of Selegiline on speech performance in *Parkinson's disease*. Folia Phoniatrica, **45**, 40-46, 1993.

[4] Teston, B. Evaluation acoustique des dysarthries : méthodes acoustiques et aérodynamiques, in Auzou, P., Ozcancack, C., Brun, V., (Eds.), *Les dysarthries, (Problèmes en médecine de rééducation, 41)*, Masson, Paris, pp. 90-108, 2001.

[5] Lagrue, B., Mignard, P., Viallet, F., Gantcheva, R., Voice in Parkinson disease: A study of pitch, tonal range and fundamental frequency variations, *ICPhs* San Fransisco, Vol. 9, pp. 1811-1814, 1999.

[6] Gentil, M., Pollack, P., Perret, J., La dysarthrie parkinsonienne, *Revue Neurologique*, 151, n° 2, pp. 105-112, 1995.

[7] Le Dorze, G., Ryalls, J., Brassard, C., Boulanger, N., et Ratte, D. A comparison of the prosodic characteristics of the speech of people with Parkinson's disease and Friedrich's ataxia with neurologically normal speakers. *Folia Phoniatrica et Logopaedica*, **50**, 1-9, 1998.

[8] Laur, D., Vigouroux, N., Nespoulous, J. L : Les altérations de la parole dans la maladie de Parkinson : bilan et perspectives de recherche, dans *Cahiers du Centre Interdisciplinaire des Sciences du Langage n° 11*, 1995-1996, Université Toulouse-Le Mirail, pp 49-60, 1996.

[9] Hoehn, M.-M., Yahr, M.-D. Parkinsonism: onset progression and mortality, *Neurology*, **17**, 472-442, 1967.

[10] Leon PR., *Phonétisme et prononciation du français*, Paris, Nathan**.** 1992

Corrélats auditifs et cognitifs à la capacité de restauration de la parole accélérée

Caroline Jacquier et Fanny Meunier

Laboratoire Dynamique du Langage (UMR 5596 – Université Lumière Lyon 2)
14, avenue Berthelot – 69363 Lyon Cedex 07, France
Mél: *jacquier@isc.cnrs.fr, fanny.meunier@univ-lyon2.fr*

ABSTRACT

We explore the relationship between auditory measures, reading capacities and the ability to reconstruct time-compressed speech for individuals without language trouble. We focused on two short attributes of speech: Voice Onset Time (VOT) and second formant transition. Normal hearing subjects had to identify disyllabic CVCV non-words that have been time-compressed on both acoustic cues simultaneously. The time compression experience showed a large inter-individual variance and allowed us to contrast a good performer and a bad performer groups for speech perception. Complementary studies (audiometric test and reading skills evaluation) showed that there is no correlation between auditory measures and cognitive mechanisms of degraded speech reconstruction whereas there are specifics correlations between reading capacities and performances in cognitive reconstruction.

1. INTRODUCTION

La compréhension du langage parlé est une tâche complexe menée quotidiennement qui représente un haut degré d'implication des fonctions cognitives. L'étude des enfants ayants des troubles du langage et de l'apprentissage -comme les dyslexiques qui ont un problème d'apprentissage de la lecture sans déficit intellectuel ou troubles neurologiques- révèlent des difficultés à percevoir des segments brefs du signal de parole [1]. Selon l'hypothèse auditive de la dyslexie, il existerait une corrélation entre la perception auditive et les capacités de lecture. Tallal a montré, par exemple, qu'il existait une grande variabilité dans la capacité à lire des non-mots chez des sujets dyslexiques. Ce type de résultats suggère que le dysfonctionnement perceptif auditif affecterait les capacités à utiliser les compétences phoniques indispensables à la lecture. Mais, d'autres études ont montré des résultats contradictoires. Les déficits des mauvais lecteurs seraient imputables à leur perception spécifique de la parole c'est-à-dire à sa nature phonologique et non auditive. Le but de notre étude est d'établir pour des sujets sans trouble du langage les relations entre performances auditives, performances phonologiques et performances de lecture.

La variabilité intrinsèque du signal acoustique de la parole pose un important problème pour la modélisation de la compréhension de la parole. En effet, le signal de parole est constitué de nombreux segments acoustiques modulables qui ont des degrés d'importance différents dans la perception de la parole [2-4]. Les recherches sur l'hypothèse phonologique de la dyslexie mettent eux en évidence des problèmes d'intégration, de traitement des sons rapides et brefs de la parole. Des déficits cognitifs dans la représentation phonologique des mots induiraient une faible conscience phonémique et des problèmes de traitement segmental des stimuli de parole.
Dans notre étude nous nous sommes intéressés à deux indices acoustiques rapides et brefs : le Délai d'Etablissement du Voisement (DEV) et la Transition du Formant 2 (TF2).

1.1. Les indices acoustiques

Le Délai d'Etablissement du Voisement (DEV)

Selon Lisker et Abramson [5], le DEV est défini comme l'intervalle de temps entre l'explosion de l'occlusive et le début du voisement. Le DEV peut-être négatif si le voisement débute avant la fin de l'explosion, nul si la synchronisation est parfaite et positif si le voisement commence un certain temps après la fin de l'explosion. L'aspiration correspond à un souffle s'échappant des poumons après l'explosion, ce phénomène est mis en évidence par un DEV positif plus long. Ainsi, les valeurs de DEV rendent compte du degré de voisement des consonnes.

La Transition Formantique (TF)

La TF correspond à un changement rapide de fréquence au moment de l'explosion de la consonne occlusive. Les changements rapides de fréquence sont primordiaux pour l'identification des segments acoustiques. La transition du second formant est un indice pour déterminer le lieu d'articulation des occlusives.

1.2. La reconstruction de la parole

La compréhension de la parole chez les sujets normo-entendants est une faculté cognitive très robuste qui résiste aux variabilités acoustiques du signal de parole. Certains mécanismes cognitifs semblent compenser et permettre la reconstruction du signal de parole altérée [6]. Ces capacités de reconstruction dépendent à la fois de la nature et du degré de distorsion appliqués au signal mais aussi de chaque individu : les capacités de perception et de compréhension de la parole dégradée sont propres à chacun, et une grande variabilité inter-individuelle est parfois observée dans des tâches d'intelligibilité.
Le but de notre étude est, d'une part, d'observer cette variabilité de reconstruction cognitive inter-individuelle à

partir d'une dégradation spécifique du signal de parole, et d'autre part, de tenter de comprendre et de circonscrire cette variabilité inter-individuelle chez des sujets normo-entendants et sans trouble du langage. Pour dégrader le signal, nous nous sommes intéressés à la dimension temporelle du signal acoustique en accélérant certains segments du signal. Ce paramètre semblant être critique pour la population dyslexique, nous avons voulu savoir ce qu'il en était chez les normo-entendants et si une grande variabilité était observable pour cette population. Nous avons manipulé deux indices acoustiques : le Délai d'Etablissement du Voisement (DEV) et la Transition du Formant 2 (TF2). Nous avons donc mesuré les effets de la compression temporelle d'indices acoustiques sur l'intelligibilité de la parole sur des sujets sans trouble du langage. Une expérience a été effectuée dans laquelle nous avons accéléré les deux indices simultanément. Nous avons ensuite conduit des études complémentaires afin de voir si les variabilités de performances observées étaient liées à des caractéristiques audiologiques et/ou cognitives de lecture propres à chacun des participants.

2. EXPÉRIENCE de COMPRESSION TEMPORELLE

Trente deux participants, âgés de 18 à 32 ans, de langue maternelle française, et n'ayant jamais connu aucun trouble auditif, du langage, ou neurologique, ont participé à cette expérience.

2.1. Matériel et Méthode

Les stimuli étaient composés de 64 non-mots bisyllabiques de forme CVCV et de 16 fillers de forme VCV. Quatre consonnes occlusives (/b/, /d/, /p/, /t/) et deux voyelles (/a/ et /i/) ont été combinées pour construire chaque stimulus. Chaque consonne apparaît avec chacune des autres consonnes dans les deux syllabes et avec les deux voyelles ($4C_1 x 4C_2 x 2V_1 x 2V_2$ = 64 CVCV). Les stimuli ont été produits par un locuteur français et enregistrés dans une chambre sourde. Les fichiers sons étaient sauvegardés sous le format wav et échantillonnés à 22 kHz (stéréo, 16 bits). La durée de chaque indice acoustique (le DEV et la TF2) a été mesurée manuellement pour chacun des items à l'aide du logiciel Praat. De même, le DEV a été segmenté manuellement à partir du début des pulsations périodiques régulières et de la détente de l'occlusion : le DEV est positif pour les occlusives non voisées et négatif pour les voisées. La TF2 a été délimitée à partir du changement brusque et rapide de la fréquence du F2, lors de la transition entre la consonne et la voyelle, jusqu'à la partie stable de la voyelle mesurée sur une représentation spectrographique en délimitant la zone où les formants sont parallèles à l'axe du temps. Pour les deux indices de chaque syllabe, la durée a été accélérée selon quatre conditions expérimentales de compression : une condition contrôle correspondant à la durée initiale, une condition 50% correspondant à 50% de la durée initiale, de même, une condition 25% et une condition 0% où les deux indices sont entièrement supprimés. La compression temporelle des indices acoustiques se fait par la méthode PSOLA (Pitch-Synchronous Overlap Add). Le but de l'analyse est d'effectuer un fenêtrage exactement centré sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition-addition (overlap-add) de ces formes d'onde élémentaires. De plus, les parties segmentées du signal acoustique (le DEV et la TF2) peuvent être accélérées alors que le reste du signal reste intact.

Les participants étaient assis dans une pièce silencieuse face à un écran d'ordinateur. Les stimuli étaient émis en modalité auditive binaurale à l'aide d'un casque (Beyerdynamic DT 48, 200Ω) et présentés dans un ordre aléatoire différent pour tous les participants. Les participants étaient informés qu'un signal de parole, pas nécessairement un mot, allait être présenté dans le casque et qu'ils devaient taper sur le clavier ce qu'ils avaient entendu. Un entraînement leur était auparavant proposé.

2.2. Résultats

Nous avons calculé les pourcentages d'identification des participants pour les items, les consonnes et les voyelles. Globalement, les voyelles étaient mieux identifiées que les consonnes. On observe pour les consonnes un effet de toutes les conditions de compression [$F(3,93)=652.32$, $p<.05$]. Les performances d'identification des consonnes dépendent aussi de la position de celles-ci : la première consonne était moins bien identifiée (70,4%) que la seconde (77%) [$F(1,31)=28.97$, $p<.05$]. On observe également une interaction entre ces deux facteurs [$F(3,93)=5.75$, $p<.05$] : selon la position (première ou deuxième syllabe), l'intelligibilité de la consonne était différemment modulée par la compression. Par ailleurs, nous avons observé une grande variabilité inter-individuelle des performances. Cette variabilité entre les 32 participants est plus importante pour la condition 25% pour les deux consonnes (SD=0.13 et SD=0.15). Par exemple, pour C1, les performances vont de 94% pour le meilleur sujet à 25% d'identification pour le moins bon. Cette variabilité reflète une différence dans les capacités à reconstruire la parole dégradée.

3. ETUDES COMPLEMENTAIRES

Suite aux résultats de l'expérience, qui montrent une grande variabilité des performances entre les participants, nous avons formé deux groupes de 12 sujets correspondant aux sujets les plus extrêmes dans leur performance à reconstruire la parole dégradée (groupes : Haute Performance, HP et Basse Performance, BP). Chaque sujet a passé un audiogramme et un test d'évaluation des capacités cognitives de lecture. Ces deux tests ont pour objectifs, d'une part, de valider la normalité des capacités auditives et de lecture des sujets et d'autre part, de mettre en évidence d'éventuelles corrélations entre les performances comportementales observées chez nos sujets et leurs capacités auditives et leurs capacités de lecture. En effet, les différences entre les sujets pourraient être liées autant à des déficits de langage (de type 'dyslexie') qu'à des troubles auditifs (perte de certaines fréquences acoustiques). Existe-t-il un lien entre l'audition et les performances cognitives de restauration de la parole ou/et un lien entre la lecture et les performances cognitives de reconstruction de la parole ?

3.1. Test d'audiométrie tonale

L'audiométrie tonale consiste à déterminer, pour plusieurs fréquences connues (125 à 8000 Hz), les seuils d'audition subjectifs d'un sujet. Le spectre de la voix s'étendant de 500 Hz à 2 kHz (voyelles : basses fréquences, consonnes : hautes fréquences). Les tests ont été pratiqués au laboratoire dans un caisson d'enregistrement insonorisé afin de s'isoler des bruits ambiants. L'audiomètre génère des fréquences à différentes intensités. Les sons purs sont présentés séparément pour chacune des oreilles à l'aide d'un casque. Leur intensité diminue progressivement jusqu'à ce que le sujet ne puisse plus l'entendre : c'est le seuil d'audition.

3.2. Test de lecture

Pour l'évaluation des capacités de lecture, nous avons utilisé le logiciel ECCLA (Evaluation diagnostic des Capacités Cognitives du Lecteur Adulte) conçu par Zagar, Jourdain et Lété, en 1995 [7]. Ce logiciel permet de déterminer la ou les sources de difficultés d'un lecteur. En effet, la lecture est conçue comme une activité cognitive modélisée par une suite d'étapes de traitement : visuel, phonologique et lexical. Le test est composé de 15 épreuves, divisé en 4 passations durant lesquelles nous enregistrons les temps de réaction des sujets. Différentes épreuves permettent l'étude de chaque étape de traitement. Pour l'étape de traitement visuel, les épreuves portent sur le codage des lettres, le jugement d'identité de lettres et le jugement de similitude de deux ensembles de lettres, prononçables ou non. Pour l'étude de l'étape de traitement phonologique, les épreuves sont une tâche de décision phonologique à choix forcé et une tâche de décision lexicale phonologique. De nombreuses épreuves évaluent l'étape de traitement lexicale : une tâche d'amorçage orthographique, des tâches de décision lexicale avec des effets de fréquence, de régularité et de la classe grammaticale, une tâche de décision phonologique à choix forcé et une tâche de catégorisation sémantique. De plus, les caractéristiques de la lecture (vitesse, temps de lecture, effet de fréquence, effet du contexte…) sont étudiées lors d'une tâche de lecture de textes mot à mot et une tâche de lecture de mots hors contexte. Ce test étant relativement long (environ 1h45), nous l'avons coupé en deux parties, de deux passations chacune, séparées par le test d'audiométrie qui durait 15 minutes.

3.3. Résultats

Audiométrie Tonale

Nous observons une diminution du seuil auditif pour les hautes fréquences (à partir de 4000 Hz) en particulier pour 6000 Hz. En effet, à 4000 et à 6000 Hz pour l'oreille droite, le groupe BP a un seuil auditif significativement plus bas (4000 Hz=5 dB, 6000 Hz=10 dB) que le groupe HP (4000 Hz=1.67 dB, 6000 Hz=4.58 dB) ainsi qu'à 6000 Hz pour l'oreille gauche (BP=10.83 dB, HP=4.58 dB).

Capacités de lecture

Le test ECCLA a fourni une importante masse de données qui ont été analysées de manière exhaustive. Nous ne présenterons ici que les résultats les plus pertinents à notre étude. Le test de Student, qui compare les moyennes des groupes pour chaque indicateur, a révélé des différences significatives pour deux indicateurs seulement, un pour l'étape visuelle et un pour l'étape de lecture. Le premier indicateur « visuel » montre que la reconnaissance des mots semblent être plus efficace pour le groupe HP (6.67% d'erreurs) que pour le groupe BP (13.33%) lorsque nous leur demandons de juger de la similitude de deux groupes de lettres non prononçables qui diffèrent uniquement par une seule lettre (Ex. BCDTF BCDTG). Ce résultat est intéressant car cette différence entre les deux groupes n'est pas observée dans le cas où les suites de lettres sont prononçables. Le second indicateur « lecture » met en évidence un effet de fréquence lors de la lecture de mots hors contexte, c'est-à-dire que les temps de lecture des sujets BP sont plus longs que les sujets HP pour les mots de basse fréquence. Il semble donc que globalement le groupe HP soit de très bons lecteurs.

Corrélations : Reconstruction/Lecture pour HP et BP

Les résultats du groupe HP montrent uniquement des corrélations entre les performances et des indicateurs de l'étape de lecture. On observe pour ce groupe une corrélation négative entre les performances de reconstruction et le style de lecture ($r=-0.58$, $p<.05$), c'est-à-dire que plus les sujets HP lisent vite meilleurs ils sont à reconstruire la parole dégradée. Cette corrélation négative se retrouve entre les performances de reconstruction et le temps de lecture de mots fréquents ($r=-0.62$, $p<.05$).

Les résultats du groupe BP montrent de nombreuses corrélations entre leurs performances et des indicateurs des étapes de reconnaissance des mots. Une corrélation est observée pour une tâche de jugement de similitude de 2 suites de lettres prononçables, en effet, la diminution des performances de reconstruction est liée à une augmentation des temps de réponse ($r=-0.82$, $p<.01$). De même, dans une tâche de décision phonologique, la diminution des performances de reconstruction est corrélée à une augmentation des temps de réponse ($r=-0.60$, $p<.05$). Enfin, un résultat très intéressant est la corrélation négative observée dans une tâche de décision lexicale : la diminution des performances de reconstruction est liée à une augmentation des temps de réponse pour les mots irréguliers ($r=-0.67$, $p<.05$). Certains indicateurs portant sur les capacités propres à la lecture corrèlent également significativement avec les résultats de reconstruction de la parole dégradée des sujets. Ainsi les temps recueillis dans une tâche de lecture de mots hors contexte corrèlent négativement avec les performances de reconstruction ($r=-0.58$, $p<.05$). On observe également une corrélation entre la sensibilité à la longueur des mots dans une tâche de lecture de textes mot à mot et les performances de reconstruction ($r=-0.63$, $p<.05$) : plus les participants sont sensibles à la longueur des mots, moins ils sont performants dans la reconstruction du signal de parole dégradée. Enfin on observe une corrélation entre l' effet de contexte phrastique sur les temps de lecture et les performances de reconstruction ($r=0.69$, $p<.05$).

4. DISCUSSION

Dans cette étude, nous avons observé, dans une première partie, les effets de la compression temporelle de deux indices acoustiques brefs (le DEV et la TF2) sur l'intelligibilité de la parole par des sujets normo-entendants. Nous avons retrouvé une variabilité inter-individuelle des performances pour la perception auditive de non-mots, surtout à la condition de compression 25%. Cette variabilité pourrait être liée à un déficit auditif ou à un trouble d'apprentissage de la lecture comme Tallal l'a observé chez les dyslexiques. Parmi les sujets, nous avons distingué deux groupes : un groupe de personnes avec de hautes performances (HP) et un groupe avec de basses performances (BP). Pour tenter de comprendre et de circonscrire cette variabilité nous avons testé la perception auditive et les capacités de lecture des deux groupes.

Les résultats des tests audiométriques rendent compte d'une bonne audition de l'ensemble des sujets même si nous avons observé une augmentation des seuils auditifs pour les hautes fréquences, à partir de 4000 Hz. Ce résultat correspond aux travaux établissant une perte à 6000 Hz comme le premier indice de la perte auditive induite par le bruit. Dans notre étude, nous avons observé une légère asymétrie et une prédominance à droite de cette perte des hautes fréquences chez le groupe BP. L'asymétrie à droite pourrait expliquer la diminution des performances chez le groupe BP. En effet, les voies auditives afférentes se croisent au niveau de l'olive supérieure donc les informations auditives captées par l'oreille droite vont être projetées sur les aires auditives primaires du cortex gauche. Les aires du langage se trouvent également principalement dans l'hémisphère gauche donc la perte d'information au niveau de l'oreille interne droite va se répercuter sur le traitement du langage. Par ailleurs, au niveau acoustique, la bande de fréquence d'une consonne se situe entre 1500 et 6000 Hz alors que celle d'une voyelle est entre 250 à 1000 Hz. Donc, l'augmentation des seuils auditifs à partir de 4000 Hz pourrait expliquer, de manière générale, le déficit d'identification des consonnes par rapport aux voyelles. Par conséquent, il sera plus difficile de percevoir ces sons consonantiques et donc de reconnaître les stimuli de forme CVCV.

Les résultats du test d'évaluation des capacités de lecture nous conduisent à de multiples discussions. Tout d'abord, suite aux analyses statistiques nous avons observé des effets de prononçabilité et de longueur des non-mots sur les temps de réponse des sujets. Une suite de lettres non prononçable est très rapidement traitée par le groupe HP alors que les non-mots de deux syllabes sont traités plus lentement par le groupe BP. Pris ensemble, ces résultats nous permettent de suggérer que les difficultés du groupe BP pourraient être dues à une mauvaise activation des codes graphémiques et donc une mauvaise représentation des formes graphiques des lettres. Cette représentation graphémique entre en jeu, principalement, dans la voie phonologique de reconnaissance des mots. La voie phonologique permet de traiter les nouveaux mots ou les non-mots comme ceux utilisés dans notre expérience de compression temporelle. Donc, si la représentation graphémique de cette voie est déficiente, l'étape de transformation des codes graphémiques en codes phonologiques serait perturbée et l'activation du lexique par le biais des codes phonologiques ou directement par les codes graphémiques serait plus faible. La reconnaissance des mots serait ainsi altérée. Ensuite, lors de l'analyse des corrélations, nous avons remarqué que les deux groupes n'évoluaient pas dans le même sens selon les étapes de traitement. Il apparaît que le groupe HP soit composé de très bons lecteurs alors que le groupe BP regroupe des sujets ayant des problèmes dans la reconnaissance des mots. Les résultats des corrélations permettent d'appuyer l'hypothèse émise précédemment sur les causes probables des déficits cognitifs des sujets BP. Nous avons montré, chez les sujets BP, que l'activation des chaînes de caractères et de codage de la position des lettres était déficiente dans l'étape visuelle. De même, nous avons mis en évidence des difficultés dans la transformation des codes graphémiques en codes phonologiques, ainsi que dans l'accès au lexique, c'est-à-dire lors de l'appariement du pattern graphémique avec une entrée du lexique mental du sujet. La reconnaissance des mots irréguliers nécessite un accès au lexique par la voie directe, l'utilisation par le système de correspondance graphème-phonème conduisant à une réponse erronée.

5. CONCLUSION

Une faible relation entre la perception auditive et la reconstruction de la parole dégradée a été mise en évidence pour l'atteinte des hautes fréquences. Par contre, il semblerait que les étapes cognitives utilisées lors de la lecture soient déficientes chez les personnes qui ont des difficultés à reconstruire de la parole accélérée.

BIBLIOGRAPHIE

[1] P. Tallal. Auditory temporal perception, phonics, and reading disabilities in children, *Brain and language*, 9:182-198, 1980.

[2] W. Serniclaes. Etude expérimentale de la perception du trait de voisement des occlusives du Français. *Ph. D. Dissertation, Université Libre de Bruxelles, 1987.*

[3] R. D. Kent and K. L. Moll. Vocal-tract characteristics of the stop cognates. *Journal of Acoustical Society of America,* 46:1549-1555, 1969.

[4] L. Lisker and A. S. Abramson. Some effects of context on voice onset time in English stops. *Language and Speech,* 10:1-28, 1967.

[5] L. Lisker and A. S. Abramson. A cross-language study of voicing in initial stops: acoustical measurements. *Word,* 20:384-422, 1964.

[6] F. Meunier, T. Cenier, M. Barkat, and I. Magrin-Chagnolleau. Mesure d'intelligibilité de segments de parole à l'envers en français. In *proc. of XXIVèmes Journées d'Etude sur la Parole,* pages 117-120, 2002.

[7] D. Zagar, C. Jourdain and B. Lété. Le diagnostic cognitif des capacités de lecture : le logiciel ECCLA (Evaluation diagnostic des Capacités Cognitives du Lecteur Adulte). *Revue Française de Pédagogie,* 113:19-29, 1995.

Familiarité aux accents régionaux et identification de mots

Girard, Frédérique (1), Floccia, Caroline (1), Goslin, Jeremy (2)

(1) Laboratoire de psychologie EA3188, Université de Franche-Comté,
30, rue Mégevand – 25000 Besançon
tel: (33) 3 81 66 59 20
fax: (33)3 81 66 54 40
mail : frederique.girard@univ-fcomte.fr
(2) School of Psychology, University of Plymouth, UK

ABSTRACT

In this study on regional accent perception we conducted two experiments to examine the role of familiarity with a given regional accent upon the observation of a word identification cost. Participants were asked for a lexical decision on the last item of sentences uttered in a familiar or an unfamiliar regional accent. In the first experiment, a group of Besançon participants were presented with their home accent and a Toulouse accent. In the second experiment, a group of Toulouse participants were presented with the same stimuli, as well as with a list of Swiss French accented sentences. Results showed an interaction between participant groups and accent familiarity, suggesting that the word identification cost associated with a non-native regional accent can be predicted by participants' familiarity with this accent.

1. INTRODUCTION

Les variabilités inter et intra-locuteur sont au cœur des recherches traitant de la description ou du traitement du signal de parole. Les sciences cognitives portent un intérêt croissant à la nature des mécanismes de normalisation et d'adaptation à cette variabilité (Lautrey, Mazoyer & van Geert, [1]). Une forme de variabilité dans la parole peu étudiée sous l'angle psycholinguistique concerne les accents régionaux. Cette variabilité peut être définie en termes phonologiques et prosodiques.

Traditionnellement, les deux sources de variation les plus largement examinées sont la voix du locuteur (Jusczyk & Luce [2] ; Bradlow, Nygaard & Pisoni [3]) et le débit de parole (Dupoux & Green [4] ; Sebastián-Gallés et al. [5]). Un point de départ possible pour l'étude de la perception des accents est l'analyse des effets produits par l'adaptation à la parole comprimée, une autre forme de variabilité inter-locuteur. Les variations de débit mènent à la démonstration de phénomènes très locaux d'adaptation tels que la normalisation de traits phonétiques temporels tels que le VOT (Miller et Liberman [6]). Face à des modifications extrêmes du débit de parole, on observe un coût sur la reconnaissance de mots. Ce coût peut s'adapter après 5 à 10 phrases, et ce type d'adaptation semble résulter de l'action conjuguée de deux mécanismes, d'un ajustement à court terme aux paramètres locaux, et d'un apprentissage à long terme qui code l'information phonologique et lexicale sur ce nouveau modèle de la parole (Dupoux & Green [4]). Ce mécanisme d'adaptation, particulièrement robuste, peut se transférer à travers un changement de locuteur ou même à travers un changement de langue, à condition que les deux langues partagent certaines caractéristiques rythmiques.

Similairement, nous avons montré au cours des années précédentes (Floccia et al [7], Girard et al. [8] [17]) que la présentation d'un accent régional non familier produit un coût initial de reconnaissance des mots, qui nécessite une certaine portion de signal pour s'établir, et qui s'habitue dans les trois premières phrases (Floccia et al. [9]). Nous présentons ici deux expériences conçues pour déterminer le rôle de la familiarité à un accent régional dans ce processus de normalisation. Nous avons employé une tâche de décision lexicale de mot cible (ou de pseudo mot) placé en fin des phrases prononcées par des locutrices avec différents accents régionaux.

Les analyses phonologiques et phonétiques du français suggèrent une large distinction perceptive entre le nord, comprenant l'accent parisien standard (accent entendu dans les médias) et des accents méridionaux (Carton, Rossi, Autesserre et Leon [10] ; Hintze, Pooley & Juge [11]). Ceci correspond également à la frontière entre les langues d'Oïl dans le nord, est, ouest ainsi que le massif central, et les langues d'Oc, qui couvrent la région du sud de la France. Une troisième famille est identifiée comme langue franco-provençale, qui englobe les zones géographiques comprenant Grenoble, Lyon et Genève, ainsi que la région Suisse Romande de la Suisse et la vallée d'Aoste en Italie (Battye, Hintze et Rowlett [12]; Rash [13], Singy [14]). Dans cette étude, nous avons contrasté un parler d'Oïl (Besançon), un parler d'Oc (Toulouse), et un parler franco-provençal (Fribourg).

Dans l'Expérience 1, notre but était simplement de répliquer un résultat antérieur (Girard et al. [15]), à savoir qu'un accent régional a priori non familier (Toulouse) engendrerait un coût de traitement par rapport à l'accent natif des participants (Besançon). L'Expérience 2 testera une population de la région toulousaine avec le même matériel, pour tenter de mettre en évidence un pattern de résultats inverse. Nous tenterons ainsi de montrer que

c'est la familiarité avec un accent régional qui est à l'origine d'un traitement plus ou moins rapide des mots, et non les caractéristiques particulières d'un accent.

2. EXPERIENCE 1

Cette expérience a pour but de mettre en évidence un coût perceptif inhérent à la présence d'un accent régional non familier.

2.1. Méthode

Participants : Trente cinq sujets monolingues francophones (6 hommes et 29 femmes), âgés en moyenne de 21;5 ans (de 18 à 37 ans). Les sujets sélectionnés étaient tous franc-comtois, sélectionnés d'après un questionnaire sur leur histoire linguistique

Stimuli : Dix mots et six pseudo-mots bisyllabiques de haute fréquence (pour les mots) ont été placés à la fin de phrases porteuses. Chaque mot ou pseudo-mot apparaît dans 4 phrases différentes, réparties entre deux accents (familier et non familier), et deux locutrices par accent (quatre locutrices naïves ont produit chacune une partie des stimuli : deux locutrices franc-comtoises et deux locutrices toulousaines). Soixante-quatre phrases de 17 à 19 syllabes ont ainsi été construites, comportant le mot-cible en dernière position. Chaque phrase était conçue de manière à ne pas permettre de deviner l'identité du mot cible, ce qui a été contrôlé grâce à des tests de prédictibilité dans une étude pilote. Après une phase d'entraînement avec des phrases et des mots/pseudo-mots n'appartenant pas au corpus de test, le bloc de 64 phrases était présenté. L'ordre de présentation des phrases était randomisé à l'intérieur du bloc pour chaque sujet. La tâche était une décision lexicale sur le dernier item de chaque phrase, aucune réponse n'était attendue sur les pseudo-mots. Les temps de réaction étaient mesurés à partir du début de la présentation du mot-cible.

2.2. Résultats

Sur les 1300 réponses obtenues, on répertorie 0.5% d'erreurs, et 2.6% de réponses lentes (supérieures à 2.5 ET de la moyenne de chaque sujet). En ce qui concerne les pseudo-mots, 2.8% de fausses alarmes sur 840 réponses ont été rapportées.

Les résultats montrent tout d'abord un effet principal de l'accent par sujet: les sujets sont plus rapides pour identifier les mots-cibles avec l'accent natif que les stimuli produits avec l'accent non-natif (545 ms versus 572 ms) ($F1 (1,34) = 23,26$ $p < .001$; $F2 (1,9) = 7,33$ $p = 0,024$), répliquant ainsi l'effet observé par Floccia et al. [7]. On observe aussi qu'il n'y a pas d'effet de la locutrice sur l'accent familier ($F1 (1, 34) < 1$; $F2 (1, 9) < 1$), mais il y en a un sur l'accent non familier ($F1 (1, 34) = 16,1$, $p < .001$; $F2 (1, 9) = 2,27$, $p = .17$).

Un examen des corrélations entre longueur des stimuli (phrases ou mots cibles) et temps de réaction a permis dans un deuxième temps d'examiner si les résultats ne sont pas dus à de simples différences de durées entre accents. La corrélation entre la longueur des phrases et les temps de réaction est significative (coefficient de régression moyen: 0,033, $t (34) = 2,95$, $p = .006$), ainsi que la corrélation entre la durée du mot-cible et les temps de réaction (coefficient de régression moyen: 0,302, $t (34) = 8,38$, $p < .001$) Ainsi, plus le mot cible et la phrase porteuse sont longs, plus la réponse est lente.

Table 1 : TR moyen et durée des stimuli

Accent		Durée des phrases (jusqu'au mot cible)	Durée mots cibles (ms)	TR moyen (ms)
Familier	loc 1	2470	492	545
	loc 2	1996	492	544
Non familier	loc 1	2382	531	598
	loc 2	2099	518	554

2.3. Discussion

Le coût de traitement induit par un accent régional non natif, aussi robuste soit-il, pourrait être attribué à des caractéristiques de durée des stimuli. En effet, les locutrices franc-comtoises ont produit des stimuli cibles plus courts que les locutrices toulousaines, et les temps de réaction sont plus lents pour les productions non natives. Afin de déterminer si ce sont des différences de durée qui sont à l'origine de l'effet de l'accent non natif, nous avions deux possibilités: la première était de manipuler la durée des stimuli afin de les égaliser, et tester un nouveau groupe de franc-comtois en faisant l'hypothèse que l'effet sera reproduit dans ces conditions. La seconde possibilité était de tester un groupe de toulousains avec le même matériel. Si les différences de durée seulement sont à l'origine de nos effets, les toulousains devraient manifester le même comportement que les franc-comtois, à savoir des temps de réaction plus lents pour les mots les plus longs, à savoir les productions toulousaines. Si par contre la durée n'est pas le seul facteur responsable des effets d'accent, et si c'est la familiarité avec l'un des deux accents qui produit des temps d'identification plus rapides, on devrait obtenir le pattern de résultats inverses, à savoir des temps de décision lexicale plus rapides avec les productions toulousaines qu'avec les productions bisontines.

3. EXPÉRIENCE 2

Cette expérience a été conçue pour déterminer si les effets de l'accent obtenus précédemment étaient dus à la familiarité des participants avec l'un ou l'autre des accents, ou à une difficulté spécifique du traitement de

l'accent non familier toulousain (ici, les durées plus longues).

3.1. Méthode

Participants : Dix-neuf participants (6 hommes et 13 femmes) avec un âge moyen de 23;5 ans. Tous les participants ont été testés dans la région de Toulouse et choisis en utilisant les mêmes critères que précédemment (leur région et accent familier devant être Toulouse bien entendu).

Stimuli : Les stimuli sont identiques à l'expérience 1. L'accent familier devient ainsi l'accent toulousain et l'accent non familier l'accent Franc-comtois. Parce que nous anticipions que l'accent franc-comtois de nos locutrices pourrait être confondu avec l'accent parisien, dont tous les français ont une grande habitude, nous avons introduit un troisième accent régional, supposé moins familier parce que plus spécifique encore que l'accent bisontin: un accent suisse-romand. Une des caractéristiques de l'accent suisse romand est une accentuation des syllabes non finales (a contrario de l'accent parisien standard). On peut aussi noter le débit relativement lent, comparativement au français. Cette caractéristique n'a jamais été étudiée sérieusement, mais cela fait partie des intuitions des linguistes suisses (Singy [16]). Trente deux phrases additionnelles ont été enregistrées par deux locutrices francophones monolingues natives de Fribourg (20 phrases se terminant par un mot, 12 par un pseudo-mot). Le bloc final de 96 phrases était présenté aux participants après une phase d'entraînement, et l'ordre de présentation des phrases était randomisé pour chaque sujet.

3.2. Résultats

Sur les 1140 réponses obtenues, on répertorie 18 erreurs, et 29 réponses lentes (supérieures à 2.5 ET de la moyenne de chaque sujet). En ce qui concerne les non mots, 34 fausses alarmes sur 684 réponses ont été rapportées (5.0%).

On observe un effet de l'accent ($F1 (2, 17) = 24{,}94$, $p < .001$; $F2 (2, 8) = 12{,}17$, $p = .004$) avec des décisions lexicales plus rapides dans l'accent familier que dans les deux accents non familiers (664.6 ms versus 690.7 ms). Cet effet est principalement dû à l'accent suisse romand (SR), avec les temps de réaction moyens de 717.9 ms, comparé à l'accent de Franche-Comté (FC) de 663.5 ms (accent familier vs accent non familier FC: $F1 (1, 18) < 1$; $F2 (1, 9) < 1$; accent familier vs accent non familier SR : $F1 (1, 19) = 48{,}19$, $p < .001$; $F2 (1, 9) = 26{,}28$, $p = .001$). Le comportement des sujets franc-comtois dans l'expérience 1 a été directement comparé à celui des sujets toulousains de l'expérience 2 en calculant une interaction entre le facteur groupe et l'accent (Franche-Comté vs. Toulouse). Les résultats étaient significatifs par sujet ($F1 (1, 52) = 6{,}68$, $p = .0126$, $F2 (1, 28) = 3{,}52$, $p = .071$), montrant ainsi qu'avec les mêmes stimuli, les deux groupes de participants ont un comportement différent. Un effet de locutrice a été observé pour l'accent natif, mais pas pour les deux accents non natifs (accent natif: $F1 (1, 18) = 6{,}88$, $p = .02$; $F2 (1, 9) = 2{,}60$; non natif FC: $F1 (1, 18) = 2{,}24$; $F2 (1, 9) < 1$; non natif SR: $F1 (1, 18) = 1{,}81$; $F2 (1, 9) < 1$).

Des analyses de régression ont montré que les durées des phrases prédisent les valeurs des temps de réaction seulement pour l'accent natif (coefficient de régression moyen = .055, $t (18) = 2{,}87$, $p = .01$), tandis que les durées des mots cibles prédisent significativement les temps de réaction pour les trois accents (accent natif: coef = .17, $t (18) = 2{,}34$, $p = .03$; non-natif FC: coef = 0,48, $t (18) = 7{,}37$, $p < .001$; non natif SR: coef = 0,40, $t (18) = 6{,}65$, $p < .001$). Ici aussi, nous trouvons que les mots les plus longs produisent les temps de réaction les plus lents. Néanmoins, alors que les stimuli toulousains sont produits avec un débit plus lent que les stimuli bisontins, les temps de réaction pour ces deux accents sont équivalents. Par ailleurs, le Tableau 2 montre que les mots produits avec l'accent suisse-romand sont plus longs que ceux produits dans l'accent natif ($t (19) = 4{,}18$, $p < .001$), ce qui pourrait expliquer pourquoi les participants étaient plus lents pour identifier ces mots. Cependant, en comparant les productions des deux locutrices fribourgeoises, on s'aperçoit que la locutrice la plus lente (643 vs. 568 ms, $t (9) = 3{,}67$, $p = .005$) est également celle qui déclenche des temps de réaction les plus rapides (708 vs. 727 ms). Ainsi, il semble peu probable que les temps de réaction plus lents obtenus avec l'accent suisse-romand soient uniquement dus à des différences de durée.

Table 2 : Durée moyenne de phrase, durée des mots, et temps de réaction (en ms) en fonction des accents et de locutrices

		Durée des phrases	Durée des mots	TR
familier	L1	2382	531	681
	L2	2099	518	648
Non familier(FC)	L1	2470	492	654
	L2	1996	492	673
Non familier(SF)	L1	3264	643	708
	L2	2931	568	727

3.3. Discussion

Dans la première expérience, les sujets franc-comtois ont montré un temps d'identification des mots plus long face à l'accent de Toulouse que face à l'accent bisontin. Dans la seconde expérience, les participants toulousains ont entendu l'accent toulousain (natif) ainsi que deux accents non natifs, l'accent franc-comtois (de Besançon) et un accent suisse romand (de Fribourg). Si les participants toulousains n'ont pas éprouvé de difficultés pour traiter l'accent bisontin par rapport à leur accent natif, ils ont en revanche été ralentis par l'accent suisse romand. Ces résultats montrent clairement que les coûts d'identification des mots lors de l'écoute d'un accent non natif sont dus au

manque de familiarité avec cet accent, plutôt qu'à certains aspects spécifiques de ce style de parole.

4. CONCLUSIONS

La distance perceptuelle entre l'accent de Franche-Comté tel qu'il est parlé à Besançon et celui du parisien, accent entendu dans tous les médias, est relativement étroite. Ceci explique sans doute pourquoi les sujets toulousains étaient relativement familiers avec ses idiosyncrasies, et n'ont pas souffert ainsi d'un retard dans la reconnaissance des mots. Il a été montré par ailleurs que les auditeurs franc-comtois se comportaient pareillement face à l'accent parisien (Floccia et al [7]). Cependant, l'accent suisse romand de Fribourg est beaucoup plus marqué et même si la plupart des français pourraient identifier, voire imiter, un accent suisse romand stéréotypé, l'exposition à cet accent reste marginale. Par exemple, l'accent suisse romand est caractérisé par une paroxytonie systématique (accentuation des syllabes non-finales) par opposition à l'oxytonie du français parisien (accentuation de la dernière syllabe). Tandis que cette propriété est également évidente dans l'accent franc-comtois elle n'est pas systématique, et est ainsi moins fréquente que dans l'accent suisse romand (Carton et al. [10]). Les résultats obtenus dans cette étude semblent bien montrer que les coûts de traitement causés par l'exposition à des accents régionaux non familiers ne sont pas spécifiques à des caractéristiques particulières de ces styles de parole, mais plutôt au degré de familiarité que les auditeurs ont avec ces accents.

BIBLIOGRAPHIE

[1] Lautrey, J., Mazoyer, B. & van Geert, P. (2002, Editeurs). Invariants et Variabilités dans les Sciences Cognitives. Editions de la Maison des Sciences de l'Homme

[2] Jusczyk, P. W. & Luce, P. A.: Speech perception and spoken word recognition : Past and present. Ear and Hearing, Vol. 23 (2002) 2-40

[3] Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B.: Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Perception & Psychophysics, Vol. 61 (1999) 206-219

[4] Dupoux, E., & Green, K.: Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology: Human Perception and Performance, Vol. 23 (1997) 914-927

[5] Sebastián-Gallés, N., Dupoux, E., Costa, A., & Mehler, J.: Adaptation to time-compressed speech: Phonological determinants. Perception and Psychophysics, Vol. 62(4) (2000) 834-842

[6] Miller, J. L., and Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*, 457-465.

[7] Floccia, C., Goslin, J. & Girard, F. (2004). Processing inter-speaker variability: the case of regional accents. Proceedings of the *Journées d'Etudes Linguistiques*, Université de Nantes, 63-66

[8] Girard, F., Floccia, C., Goslin, J., Konopczynski, G. (2005). Coping with an unfamiliar regional accent: a lexical decision study in French listeners. Proceedings of the Workshop PSP2005, Londres

[9] Floccia, C., Goslin, J. & Girard, F. (submitted). Time-course of adaptation to regional and foreign accents. Proceedings of 14th Manchester Phonology Meeting

[10] Carton, F., Rossi, M., Autesserre, D., & Léon, P. (1983). *Les Accents du Français*. Paris : Hachette, De Bouche à Oreille

[11] Hintze, M.-A., Pooley, T., & Judge, A. (2001). *French Accents: Phonological and Sociolinguistic Perspectives*, CILT - AFLS.

[12] Battye, A., Hintze, M.-A., & Rowlett, P. (2000). *The French Language Today: A Linguistic Introduction*, 2nd ed. London and New York: Routledge.

[13] Rash, F. (2002). The German-Romance language borders in Switzerland. *Journal of Multilingual and Multicultural Development, 23*, 112-136.

[14] Singy, P. (ed). (2002). Le Français Parlé dans le Domaine Francoprovençal : une Réalité Plurinationale, Bern: Peter Lang.

[15] Girard, F., Goslin, J., & Floccia, C.: Un accent régional perturbe-t-il l'accès lexical? Proceedings of the Workshop MIDL, Paris, (2004),161-166

[16] Singy, P. (ed.). (2004). Identités de Genre, Identités de Classe et Insécurité Linguistique en Suisse Romande. Bern : Peter Lang .

[17] Girard, F., Goslin, J., Floccia, C. (2005). Influence of regional accents in speech perception. Proceedings of the Workshop CONTEXT-05, Paris, july 2005

Nasalité consonantique et coarticulation : étude perceptive

Tiphaine Ouvaroff, Solange Rossato

Institut de la Communication Parlée
Université Stendhal - 1180, Avenue Centrale, BP 25, 38040 GRENOBLE CEDEX 9
Tél. +33 (0)4 76 82 43 37
Tiphaine.Ouvaroff@icp.inpg.fr, Solange.Rossato@icp.inpg.fr

ABSTRACT

This paper investigates the coarticulation of consonantal nasality from a perceptive point of view. The aim of this study is to determine in CV utterances to which extent the consonant is perceived in the following vowel and compare these perceptual boundaries between oral and nasal consonants. Results show that, although the vowel is actually nasalized (low velum and consistent nasal air flow), the listeners don't attribute the nasalization of the vowel to the presence of a nasal consonant. After the release, the nasal feature of the consonant is lost, only the place of articulation is perceived until 60 ms after the release.

1. INTRODUCTION

De nombreuses études se sont intéressées au phénomène de coarticulation de la nasalité (entre autres Krakow & Beddor [4], Krakow et al. [5], Abramson et al. [1]) : il est avéré que les voyelles en contexte de consonne nasale sont nasalisées par effet de propagation du trait nasal de la consonne. Des enregistrements réalisés au sein de l'ICP nous ont permis d'observer le phénomène de coarticulation de la nasalité d'un point de vue articulatoire et aérodynamique, et cette étude perceptive se place dans la continuité de ces travaux.

Des mesures EMA (Articulographie Electro-Magnétique) réalisées en 2002 nous ont permis d'évaluer d'un point de vue articulatoire le phénomène de coarticulation consonantique de la nasalité, grâce à des données quantitatives quant à l'abaissement vélaire (Rossato et al. [8]). Dans des séquences de type NV, nous avons pu observer une coarticulation progressive de la consonne nasale sur la voyelle orale. La figure 1 montre la trajectoire vélaire durant un segment VNV : le velum reste abaissé au même niveau que la consonne nasale durant toute la durée de la voyelle, cette hauteur étant dépendante de la hauteur de la voyelle.

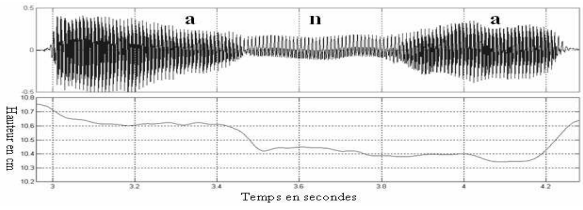


Figure 1 : Trajectoire du velum dans l'axe des Y (en bas) synchronisée au signal acoustique pour un segment [ana].

Des enregistrements aérodynamiques EVA réalisés en 2003 sur le même corpus et pour le même locuteur, ont permis l'observation d'un débit nasal important dans la voyelle suivant la consonne nasale (Ouvaroff [6]). Ce débit nasal n'est pas réservé exclusivement au [a] : les voyelles fermées [i], [y] et [u] attestent un débit nasal élevé (Figure 2), d'ailleurs supérieur à celui des voyelles nasales.

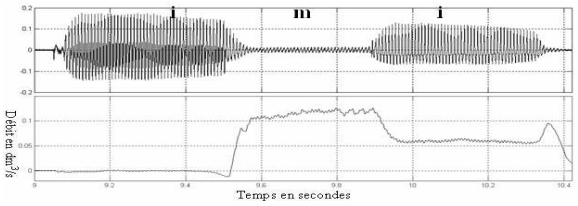


Figure 2 : Débit nasal en dm^3/s (en bas) synchronisé au signal acoustique pour un segment [imi].

Plusieurs travaux ont étudié d'un point de vue perceptif le rapport entre la consonne nasale et la voyelle en contexte. Krakow & Beddor [4] observent que la voyelle orale [ɛ], nasalisée car placée en contexte consonantique nasal [mɛn], est perçue nasale même placée en contexte CVC ou #V#. Cette voyelle est même perçue plus nasalisée qu'en contexte approprié. La nasalisation de la voyelle modifie la perception de sa hauteur : les voyelles hautes sont perçues plus basses tandis que les voyelles basses sont perçues plus hautes (Wright [11]). Krakow et al. [5] ont montré pour des locuteurs américains que ce couplage nasal n'amène pas nécessairement à une mauvaise perception de la hauteur de la voyelle quand la nasalisation de cette voyelle provient d'un phénomène de coarticulation. D'autre part, la hauteur intrinsèque du velum dépend de la hauteur de la voyelle. Abramson et al. [1] a montré que cette hauteur vélaire de la voyelle affecte les frontières perceptives de la nasalité dans un continuum acoustique [d-n]. Il existe un rapport étroit en termes de perception entre la consonne et la voyelle, et c'est sur ce rapport que s'appuie notre test de perception de la nasalité consonantique.

Il s'agit de définir, dans des séquences de type CV, les 'frontières perceptives' de la coarticulation. En Français qui oppose voyelle orale et voyelle nasale, la nasalisation par contexte de la voyelle est-elle attribuée à un phénomène de coarticulation ? Peut-on retrouver ainsi le trait nasal de la consonne précédente ? Pour cela, nous comparons les frontières perceptives de la coarticulation des consonnes [b d] versus [m n].

2. MÉTHODE

Nous avons donc réalisé un test perceptif comportant des stimuli de type 'backward gating', utilisé précédemment par Smits [9] et Pols & Schouten [7] dans leur étude sur la perception des consonnes.

2.1. Stimuli

Les signaux utilisés proviennent d'enregistrements EMA réalisés en 2002 au sein du laboratoire (dispositif EMA Carstens AG100) pour lesquels nous disposons des trajectoires des articulateurs (y compris le velum). Dans ce corpus, des séquences de la forme CV ont été sélectionnées, où C = [b d m n] et V = [a i u]. La sélection des séquences s'est faite sur la qualité acoustique du signal et sur sa régularité. Les stimuli ont été extraits par technique de gating, procédé consistant à dévoiler le signal par étapes, grâce au logiciel PRAAT. La découpe s'est faite par rapport à l'instant de relâchement acoustique de la consonne (noté '(0)'), et ce par tranches de 20 ms environ (les stimuli sont découpés au passage par zéro le plus proche). 10 stimuli ont été ainsi extraits pour chaque séquence CV : le premier a son onset 60 ms avant le relâchement de la consonne (un pré-test nous a montré que nous avons une bonne identification de al consonne) et le dernier 120 ms après le relâchement (fin des transitions formantiques). Un exemple de ce découpage est donné en Figure 3. L'offset est fixe et se situe à la fin de la partie stable de la voyelle. Nous avons des stimuli de durée variable. Notre test comporte au final (10 tranches x 4 consonnes x 3 voyelles) = 120 stimuli CV. Des stimuli de la forme VC ont également été découpés et utilisés dans le test de perception, mais sont pour l'instant en cours d'analyse.

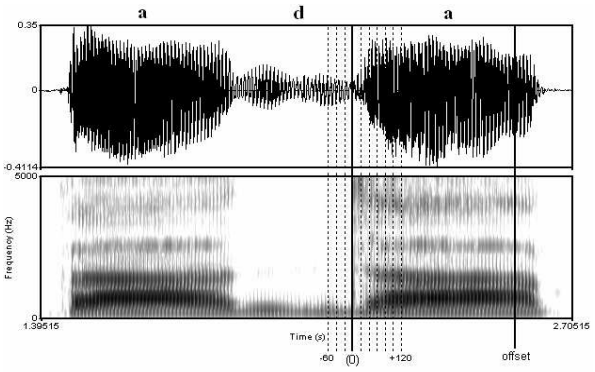


Figure 3 : Exemple de découpe des stimuli d'une séquence [ada] par technique de 'backward gating' : 10 pas de gating à offset fixe et à onset variable, de 60 ms avant jusqu'à 120 ms après le relâchement de la consonne (0).

2.2. Procédure

Le test perceptif a été implémenté grâce au logiciel Dreamcard Revolution et se divise en trois parties correspondant aux trois voyelles [a i u]. La consigne donné au sujet est la suivante : «Déterminer la consonne que vous entendez avant la voyelle, sachant qu'elle a été coupée. » Le sujet a le choix entre 5 réponses fermées : [b] [d] [m] [n] et {aucune consonne} dans le cas où la consonne n'est pas perçue par le sujet.

2.3. Sujets

16 sujets ont participé au test perceptif, mené au sein du laboratoire de l'ICP. Tous les sujets sont des français natifs, exceptés deux, de nationalité américaine et syrienne : leurs résultats ont été exploités étant donné que leur langue comporte les oppositions consonantiques labiale/dentale et orale/nasale.

3. ANALYSE DES RÉSULTATS

L'analyse des résultats des tests de perception se fait en termes de taux d'identification correcte des consonnes [b d m n], et des éventuelles différences selon le mode et le lieu de la consonne. Les taux d'identification correcte en fonction du gating sont modélisés par une régression logistique binaire 'logit' (obtenue grâce à SPSS), qui permet de déterminer la 'frontière perceptive' (seuil de 50 % de la courbe 'logit').

3.1. Identification correcte de la consonne dans la voyelle : mode et lieu

Les taux d'identification correcte des consonnes en fonction du pas de gating sont représentés Figure 4 pour les quatre consonnes.

Contraste de mode

Les courbes d'identification correcte sont significativement différentes entre les consonnes orales et nasales ($p<0,001$). Les frontières perceptives des consonnes nasales se situent au niveau du relâchement de l'occlusion (+4,8 ms pour [m], +2,2 ms pour [n]). Concernant les consonnes orales, les frontières perceptives se situent 42,2 ms après le relâchement pour le [b] et 54,3 ms pour le [d]. L'interaction entre le mode et le pas de gating n'est pas significative : les courbes d'identification correcte des consonnes orales vs nasales ont des pentes similaires et sont décalées en fonction du gating.

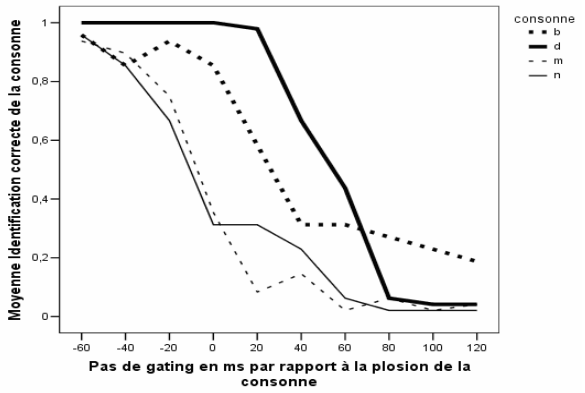


Figure 4 : Pourcentage d'identification correcte des consonnes [b d m n], en fonction du pas de gating (temps en ms par rapport au relâchement de la consonne (0)).

Nous pouvons donc dire au vu de nos résultats que les effets de coarticulation, en termes d'abaissement du velum et de débit nasal, ne permettent pas aux sujets d'attribuer la nasalisation de la voyelle à la présence d'une consonne nasale en contexte gauche.

Contraste de lieu

Le lieu d'articulation des consonnes est pertinent : les courbes d'identification correcte sont significativement différentes selon le lieu d'articulation de la consonne, et il y a une interaction significative entre le mode oral/nasal et le lieu labial/coronal ($p<0,001$). En effet, cette différence de lieu n'est significative qu'en contexte de consonne orale [b d] ($p<0,001$) : la frontière perceptive de la consonne coronale [d] est située environ 12,1 ms plus tard que celle de la consonne labiale [b]. Il existe pour les consonnes orales une interaction entre le lieu et le pas de gating : la courbe 'logit' qui modélise le taux d'identification correcte des coronales a une pente plus raide que celle des labiales. Aucun différence significative n'est trouvé entre les taux d'identification correcte du [m] et du [n] : le lieu n'est pas pertinent dans la perception de la consonne nasale précédente, contrairement au cas des consonnes orales.

3.2. Perception de la présence d'une consonne : les confusions en termes de traits d'articulation

En plus de l'identification correcte de la consonne, nous nous sommes intéressés à la perception d'une consonne quelle qu'elle soit (toutes réponses sauf [aucune consonne]). La Table 1 indique pour chaque consonne [b], [d], [m] et [n] les frontières perceptives des courbes d'identifications correctes et celles des courbes de la perception d'une consonne. Il y a un laps de temps de20 ms à 30 ms durant lequel une consonne est perçue, mais les traits de mode ou de lieu d'articulation ne sont plus correctement identifiés. Quelles sont les types de confusions opérées durant cet intervalle ?

Table 1 : Frontières d'identification correcte et de perception d'une consonne quelconque pour les consonnes [b d m n] (par rapport au burst).

	[b]	[d]	[m]	[n]
Identification correcte de la consonne	42,2 ms	54,3 ms	4,8 ms	2,2 ms
Perception d'une consonne quelconque	62,9 ms	77,9 ms	26,4 ms	35,1 ms

Notons tout d'abord que cet intervalle ne correspond pas au même pas de gating suivant la consonne : pour un pas de gating 0 ms, nous sommes proches du seuil de 50% d'identification correcte pour les consonnes nasales tandis que les consonnes orales [b] et [d] sont correctement identifiées. Ainsi, une analyse des confusions au pas de gating 0 ms permet d'observer les confusions des consonnes [m] et [n] tandis qu'une analyse au pas de gating +40 ms nous renseigne sur les confusions des consonnes orales. La Table 2 présente les consonnes perçues à ces deux pas de gating : 0 ms et +40 ms.

Nous observons les consonnes nasales que le lieu est correctement identifié, les principales erreurs concernant le mode : la consonne [m] a tendance à être perçue [b] (35,4 % des réponses), le [n] perçu [d] (39.6 % des réponses) au pas de gating 0 ms, tandis que les consonnes orales montrent peu voire pas de confusion. On observe donc principalement une confusion de mode pour les consonnes nasales.

Au pas de gating de +40 ms, les consonnes nasales ne sont plus perçues dans la majorité des cas. Pour les consonnes orales, la consonne identifiée reste le plus souvent orale (pour le [d] notamment) ou bien aucune consonne n'est perçue (52.1 % des réponses pour le [b]). Les confusions portent plutôt sur une mauvaise perception du lieu d'articulation des consonnes orales.

Table 2 : Réponses en pourcentage données au pas de gating 0 ms et +40 ms

Gating		[b]	[d]	[m]	[n]	vide
0 ms	**[b]**	**85,4**	4,2	6.2	0	4.2
	[d]	0	**100**	0	0	0
	[m]	*35.4*	2.1	**35.4**	4.1	*23*
	[n]	0	*39.6*	0	**31.2**	*29.2*
40 ms	**[b]**	**31.2**	10.4	6.3	0	*52.1*
	[d]	*27.0*	**66.7**	0	2.1	4.2
	[m]	6.2	0	**14.6**	2.1	*77.1*
	[n]	10.4	4.2	2.1	**22.9**	*60.4*

Les résultats en terme de reconnaissance du lieu sont explicables simplement par le fait que l'intervalle de confusion pour les consonnes nasales se situe en début de la transition formantique (Figure 5) : comme l'ont montré, entre autres, Stevens & Blumstein [10], les premières 26 ms de la transition formantique permettent d'identifier le lieu d'articulation de la consonne. L'intervalle de confusion des consonnes orales se situe trop loin dans la transition pour que le lieu soit encore clairement identifiable, notamment pour le [d].

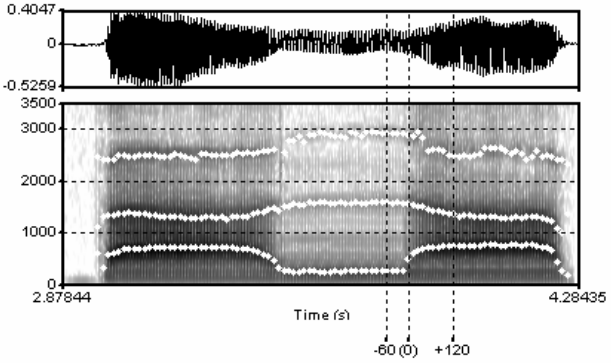


Figure 5 : Spectrogramme et contours formantiques.de la séquence [ana] avec les pas de gating de -60ms et +120ms par rapport au relâchement du [n] (0).

Un fait des plus intéressant est que les consonnes nasales, avant de ne plus être perçues du tout, sont confondues avec leur consonne homorganique orale dès la disparition du murmure nasal. La nasalisation seule de la voyelle n'est pas suffisante pour identifier la nasalité de la consonne précédente : dès que la consonne est relâchée, la nasalisation de la voyelle n'est plus attribuée à un phénomène de coarticulation. Le même type de résultat a été montré par Beddor & Onsuwan [2] sur la perception des consonnes prénasalisées. Ils observent que dans une séquence [mbV] où le [b] inséré dure environ 27 ms, même si la voyelle est nasalisée à 100% les sujets perçoivent en majorité [mb] au lieu de [m] : la nasalisation de la voyelle n'est pas suffisante pour contrer une absence de murmure nasal. Entre 0 et 20 ms après le burst, l'indice de nasalité n'est pas suffisant pour identifier clairement la consonne nasale et l'indice de présence d'une consonne – matérialisé par les transitions formantiques – indique d'après nos résultats le lieu d'articulation de la consonne : le trait de lieu est alors maintenu, le [m] peut être perçu [b] et le [n] perçu [d].

4. CONCLUSION

La nasalisation de la voyelle n'est pas suffisante pour identifier le trait nasal de la consonne précédente une fois le murmure nasal disparu, tandis que le lieu d'articulation, porté par les transitions formantiques, reste perçu. Nous ne pouvons pas dire au vu de ce seul test que la voyelle n'est pas perçue nasalisée, seulement que cette nasalisation n'est pas attribuée à la consonne adjacente. La perception de la nasalisation serait donc à tester indépendamment de la consonne, d'autant plus que la langue française comporte l'opposition voyelle orale/voyelle nasale. Il serait intéressant d'extraire la voyelle nasalisée et l'observer dans d'autres contextes, à l'instar de Krakow & Beddor [4] : elles notent, de la même façon que Kawasaki [3], que la perception de la nasalité de la voyelle n'est pas favorisée par la présence adjacente de la consonne nasale. Une prochaine étape sera également d'observer la perception de la nasalité en fonction de la qualité de la voyelle.

REMERCIEMENTS

Nous tenons à remercier Pierre Badin, notre locuteur pour les corpus aérodynamique et articulatoire, ainsi que Christophe Savariaux.

BIBLIOGRAPHIE

[1] A. S. Abramson, P.W. Nye, J. Henderson and C.W. Marshall. Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America*, 70:329-339, 1981.

[2] P.S. Beddor and C. Onsuwan. Perception of prenasalized stops. *Proceedings of 15th ICPhS*, Barcelona, Spain, 2003.

[3] H. Kawasaki. Phonetic explanation for phonologicals universals: the case of distinctive vowel nasalization. In *Experimental Phonology*, Edited by J.J Ohala, J.J. Jaeger, New York Academic Press, New York, NY, pages 81-103, 1986.

[4] R.A. Krakow and P.S. Beddor. Coarticulation and the perception of nasality. In *Proceedings of 12th ICPhS*, Aix-en-Provence, France, volume 4, pages 38-41, 1991.

[5] R.A. Krakow, P.S. Beddor, L.M. Goldstein and C. Fowler. Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America*, 83:1146-1158, 1988.

[6] T. Ouvaroff. Mesures aérodynamiques de la parole dans le cas des nasales. *Rapport de stage de maîtrise*, 2004.

[7] L.C.W. Pols and M.E.H. Schouten. Identification of deleted consonants. *Journal of the Acoustical Society of America*, 64:1333-1337, 1978.

[8] S. Rossato, P. Badin and F. Bouaouni. Velar movements in French: An articulatory and acoustical analysis of coarticulation. In *Proceedings of 15th ICPhS*, Barcelona, Spain, pages 3141-3144, 2003.

[9] R. Smits. Human consonant recognition for initial and final segments of VCV utterances. In *Speech Hearing and Langage*, Work in process 10, Department of Phonetics and Linguistics, University College London, UK, pages 115-136, 1998.

[10] K.N. Stevens and S.E. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64:1358-1368, 1978.

[11] J.T. Wright. Effects of vowel nasalization on the perception of vowel height. In *Nasalfest : Papers from a Symposium on Nasals and Nasalization*. Edited by C.A. Ferguson, L.M. Hyman and J.J. Ohala, Langage Universals Project, Stanford University, Stanford, CA, pages 373-388, 1975.

Reconnaissance automatique de phonèmes guidée par les syllabes

Olivier Le Blouch, Patrice Collen

Laboratoire TECH/IRIS
France Télécom R&D – 4 rue du Clos Courtel – 35510 Cesson Sévigné, France
Tél.: ++33 (0)2 99 12 48 51 – Fax: ++33 (0)2 99 12 40 98
Mail: olivier.leblouch@francetelecom.com & patrice.collen@francetelecom.com

ABSTRACT

This paper presents a phonetic transcription system of French speech. This recognizer is based on a phonetic transcription driven by syllables, a syllabic bigram language modelling and a HMM topology adapted to syllables. The phone error rate obtained is compared to basic, usual systems at phonetic level : once the resulting syllables have been converted to phones, the phone error rate on a 12 minute-part of BREF80 corpus is as low as 15.8% with 35 phones.

1. INTRODUCTION

La syllabe joue un rôle important dans la perception de la parole et son utilisation comme entité acoustique constitue un avantage certain vis-à-vis d'une approche par mot ou par phonème [5]. La syllabe est en effet une entité plus longue que le phonème, plus facile à indexer manuellement, et offrant la possibilité de générer un nombre de mots illimité avec un nombre de syllabes réduit, de l'ordre de 5000 [13]. En outre, des études en psycholinguistique et en phonologie suggèrent que l'information syllabique, avec une durée moyenne de 250ms, est cruciale au niveau de la perception et de la compréhension [7].

Cet article propose d'utiliser les syllabes de façon à améliorer la transcription phonétique. Notons que, contrairement au modèle syllabique décrit dans [5], le système décrit dans ce papier est conçu comme un système de transcription phonétique dont l'enchaînement est guidé par l'information syllabique.

De façon à se positionner vis-à-vis des systèmes existants, les systèmes implémentés ont été testés sur le corpus BREF80 [8], servant de base de test notamment aux systèmes décrits dans [9] et [3]. Le premier propose un système combinant des modèles de phonèmes en contexte, un modèle de langage bigram phonétique et une adaptation au genre du locuteur; ce système offre un taux d'erreur de 21.3%. Le second, quant à lui, propose une reconnaissance associant chaînes de Markov et réseaux neuronaux, et obtient un taux d'erreur de 25%.

Après une description des ressources utilisées, et un bref état de l'art sur les techniques de syllabification, nous détaillerons l'architecture générale des systèmes phonétiques et syllabiques développés. Avant de conclure, les performances sur BREF80 ainsi que la complexité des différents systèmes seront présentées.

2. RESSOURCES AUDIOS ET TEXTUELLES

2.1. Corpus Audio

Toutes les données audio sont échantillonnées à 16kHz.

Le corpus d'apprentissage, d'une durée de 4h30, est composé à 60% de données issues du corpus BREF80 et de 40% de données audio issues de corpus France Télécom. Il est au total constitué de 274 phrases prononcées par 40 locuteurs et de 310 phrases prononcées par 53 locutrices.

Le corpus de test de 12 minutes, issu du corpus BREF80, est quant à lui composé de 24 phrases prononcées par 8 locuteurs différents (4 hommes et 4 femmes), non présents dans le corpus d'apprentissage.

Les transcriptions phonétiques associées à toutes ces données ont été réalisées automatiquement sur la base des 35 phonèmes proposée dans [6], puis vérifiées manuellement afin de refléter au mieux le contexte acoustique.

2.2. Corpus textuel

Pour la création des modèles de langage bigram phonétique et syllabique, c'est-à-dire le calcul des probabilités de transition d'une entité phonétique ou syllabique vers une autre, les données textuelles présentes dans notre corpus audio d'apprentissage ont été enrichies par divers contenus en langue française issus du Web. Au total, ce corpus représente environ 300K mots, soit environ 600K syllabes et 1300K phonèmes.

3. DU TEXTE AUX SYLLABES

La décomposition du texte en entités syllabiques requiert un outil de syllabification. Pour le français, nous nous sommes inspirés des principes utilisés pour la syllabification des bases lexicales *Brulex* et *Lexique* [11].

Dans la suite de cet article, le formalisme de représentation des syllabes est un agglomérat de

phonèmes séparés par des "_", un mot de plusieurs syllabes étant quant à lui une suite de syllabes séparées par des blancs. Ex : *Syllabe* ➔ [S_I L_A_B].

Le découpage du français en syllabes obéit à certaines conventions [11]. Tout d'abord, chaque son vocalique, c'est-à-dire les voyelles et les semi-consonnes suivies d'une voyelle, constitue le noyau d'une syllabe alors que deux voyelles consécutives (Ex : *Agréable* ➔ [A G_R_**EI** **A**_B_L]) appartiennent à deux syllabes différentes : elles sont dites en hiatus. Ensuite, lorsque entre deux voyelles, une seule consonne est prononcée, elle est considérée comme formant une syllabe avec la voyelle qui la suit, et ce indépendamment du découpage en mot ; la phrase *Quelle heure est-il ?* est ainsi syllabée [**K**_AI **L**_OE **R**_AI **T**_I_L]. Enfin, dans le cas de plusieurs consonnes prononcées entre deux voyelles, il existe des règles inhérentes à chaque cas, comme par exemple les agrégats occlusives-liquides (Ex : R*empli* ➔ [R_AN **P_L**_I]) ou les suites d'occlusives (Ex : *Opter* ➔ [O_**P T**_EI]).

Néanmoins, le découpage du français en syllabes se heurte à différentes théories explicitées dans [11], comme par exemple la syllabification du mot *capsule* : [K_A_**P S**_U_L] ou [K_A **P_S**_U_L] ? Dans ce type de cas, et lorsque c'est possible, l'analyse acoustique sert de support pour le choix de segmentation : sachant que les occlusives sont précédées d'une courte période de silence, le choix s'est porté sur une segmentation avant l'occlusive et donc une syllabation en [K_A **P_S**_U_L].

L'algorithme implémenté suit donc ces conventions en convertissant dans un premier temps le texte en phonèmes, en prenant soin de substituer aux ponctuations des silences, puis en faisant appel à une soixantaine d'heuristiques de découpage des chaînes de phonèmes. Ceci aboutit à une segmentation du corpus textuel en 4352 syllabes différentes.

4. DESCRIPTION DES SYSTEMES

Après une description technique des paramètres utilisés et des algorithmes communs à toutes les expériences, les 3 systèmes sont présentés.

4.1 Paramétrage et description technique

Paramètres acoustiques

Le signal audio est converti en un jeu de 39 coefficients extraits toutes les 10ms sur des segments temporels de 32ms. Les vecteurs sont constitués de 12 coefficients MFCC, de la log-énergie, et des dérivées premières et secondes. Une normalisation par la moyenne des cepstres est finalement appliquée [1].

Modélisation

Nos unités sont modélisées par des chaînes de Markov cachées [12].

Modèles de langage

Un modèle bigram est appliqué à tous les systèmes : phonétique pour le premier système et syllabique pour les deux autres. Tous deux sont appris sur le corpus textuel à partir des modules HTK [16] dédiés.

Apprentissage des HMM

L'apprentissage est réalisé sous HTK selon l'algorithme de Baum-Welch [2] sur le corpus dédié, en multipliant par deux le nombre de gaussiennes par mixture toutes les 10 itérations. Au final, chaque état contient un mélange de 32 gaussiennes.

Décodage

Pour le décodage, on utilise une formulation alternative de l'algorithme de Viterbi, le *Token Passing Model* [14].

4.2. Système phonétique de base

Ce système met en œuvre des modèles de Markov cachés à trois états avec une topologie gauche-droite.

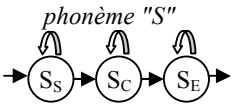


Figure 1 : Topologie des modèles de phonème

La topologie de ces modèles phonétiques utilise un formalisme simple pour décrire chaque état : pour chaque phonème X à 3 états, la notation X_S est donnée à l'état "START", X_C à l'état "CENTER" et X_E à l'état "END".

4.3 Premier système syllabique

A partir de cette base phonétique, chaque modèle de syllabe est construit par concaténation des modèles de phonèmes, comme illustré Figure 2.

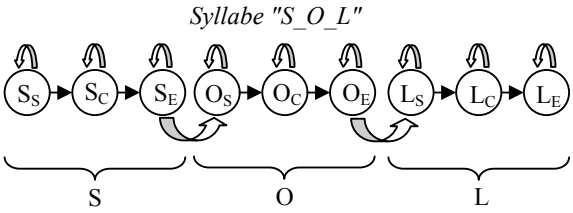


Figure 2 : Exemple de phonèmes concaténés en syllabe

Pour le décodage, on utilise le bigram syllabique décrit ci-dessus. Pour chaque trame, l'algorithme de décodage calcule ici les probabilités d'émission de 105 (35x3) états différents.

4.4 Second système syllabique

L'inconvénient majeur du système précédent est l'apprentissage de chaque modèle phonétique

indépendamment de leur contexte. Or les phonèmes, au sein d'un flux de parole, ne se suivent pas brutalement et des phénomènes de coarticulation apparaissent en fonction du contexte d'émission de chaque phonème. Il s'agit là d'un problème récurrent de la reconnaissance de parole auquel on répond le plus souvent par l'ajout d'informations contextuelles (diphones, triphones) [9].

Dans le cadre d'un système basé sur des unités syllabiques, il est souhaitable de travailler sur des modèles syllabiques cohérents à l'intérieur desquels le contexte est pris en compte et correctement modélisé. Dans cette optique, chacune des 4352 syllabes est associée à une chaîne de Markov spécifique, directement issue de la modélisation vue ci-dessus et présentée en Figure 3.

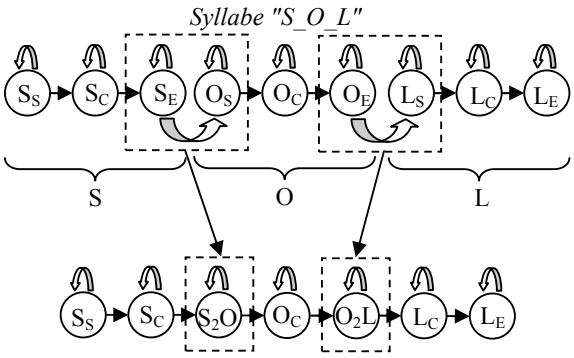


Figure 3 : Création d'un modèle syllabique

Cette modélisation conserve la topologie issue de la concaténation des chaînes de Markov phonétiques, à ceci près que les états extrémités des phonèmes contigus sont fusionnés en un seul, représentant le contexte de transition. La nouvelle notation consacrée pour ces nouveaux états transitoires entre deux phonèmes X et Y est X_2Y (X to Y). Ainsi, sur l'exemple présenté en Figure 3, les états S_E et O_S sont fusionnés en S_2O et O_E et L_S en O_2L. Cette fusion de deux états entraîne une réduction du nombre d'états à parcourir lors du décodage, car là où une concaténation de *n* phonèmes était modélisée par 3*n* états, le modèle syllabique correspondant n'en aura plus que 2*n*+1. Par ce formalisme, l'intégralité de l'espace de parole est donc couverte par un ensemble de 1295 états différents.

En outre, rappelons que ces syllabes ont été extraites d'un corpus textuel bien plus important que les seules données du corpus d'apprentissage audio. En effet, ce dernier n'en contenant que 2565, près de la moitié des modèles syllabiques ne seront pas appris directement sur le corpus audio. Malgré tout, la topologie de ces syllabes permet d'étendre l'apprentissage aux modèles non rencontrés en partageant les états [15], comme indiqué en Figure 4.

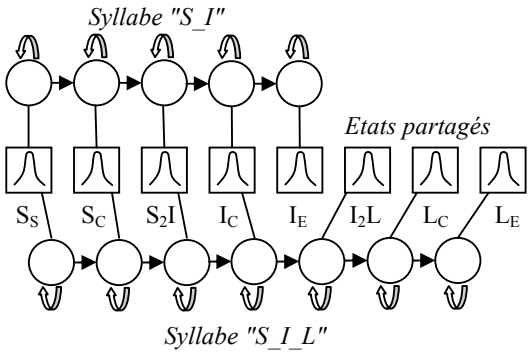


Figure 4 : Partage d'états entre syllabes

Au final, les 4352 syllabes se partagent donc 1295 états différents, c'est-à-dire près de 10 fois plus de probabilités d'émission à calculer lors du décodage que pour le système phonétique. Une fois ces prototypes générés, un apprentissage est effectué : chaque état est initialisé sur les moyenne et variance globales du corpus d'apprentissage puis ré-estimé par algorithme de Baum-Welch.

5. PERFORMANCES ET COMPLEXITES

Le Tableau 1 reprend la description technique des trois systèmes décrit précédemment.

Tableau 1 : Désignation des systèmes

Systèmes	Désignation
A	Système phonétique avec bigram phonétique (§ 4.2)
B	Système phonétique avec bigram syllabique (§ 4.3)
C	Système syllabique avec bigram syllabique (§ 4.4)

Les performances et complexité de chacun des systèmes sur le corpus de test BREF précédemment décrit sont présentées Tableau 2. Le PER (Phone Error Rate) représente le taux d'erreur par phonème:

$$PER = 100 - Accuracy = 100 * \frac{S + I + D}{N}$$

Où N correspond au nombre total de phonèmes de la transcription manuelle de référence, I aux insertions, S aux substitutions et D aux omissions. Les silences ne sont pas pris en compte dans la mesure afin de ne pas accroître artificiellement les performances [9].
Notons qu'afin de rendre comparables les performances des différents systèmes proposés, les syllabes décodées ont été retranscrites en phonèmes.
La complexité, exprimée en fonction du temps réel (RT=Real Time), est ici calculée sur un Pentium Xeon 3.7 GHz avec 2Go de RAM.

Tableau 2 : Performances des différents systèmes

Système	Corr.	Subs.	Del.	Ins.	PER	Complexité
A	74.9	16.2	8.9	1.8	**26.9**	0.11 RT
B	83.65	10.8	5.5	1.7	**18**	1.9 RT
C	86.35	9.8	3.9	2.1	**15.8**	2.3 RT

6. DISCUSSION

L'apport du bigram syllabique sur les performances est évident. En effet, le taux d'erreur est réduit de près de 9 points en construisant les syllabes à l'aide des 35 phonèmes de base et en y appliquant le modèle de langage. L'explication provient du fait que les phonèmes sont dorénavant guidés sur un sous-espace plus représentatif de la parole, chaque syllabe (concaténation de *n* phonèmes) faisant office de *n*-gram au niveau phonétique. Le système C quant à lui, en réduisant encore le taux d'erreur de près de 3 points, montre l'intérêt d'inclure des états en contexte au sein des modèles syllabiques.

Les complexités liées à l'application du bigram syllabique dépendant du nombre élevé d'unités traitées, la durée de décodage atteint ici plus de 2 fois le temps réel. Notons que la complexité est moins élevée pour le système B que pour le système C, ce que l'on peut expliquer par le nombre de probabilités d'émissions différentes à calculer pour chaque système (10 fois moins pour B que pour C).
Notons toutefois qu'avec une simple stratégie d'élagage [10] dans l'algorithme de Viterbi, le temps de décodage du système C a été ramené à une fois le temps réel pour une même performance.

7. CONCLUSION ET PERSPECTIVES

Cette utilisation des syllabes comme guides de la reconnaissance phonétique permet donc de réduire significativement le taux d'erreur sur les phonèmes en proposant un système simple et performant, fonctionnant en temps réel et conservant une bonne liberté de généralisation pour des applications de détections de mots clés ou de reconnaissance grands vocabulaires. En outre, l'ajout de l'information contextuelle au sein même des entités syllabiques accroît également les performances grâce à une modélisation plus précise du signal sur des portions plus larges.

Cependant des améliorations sont envisagées, notamment au niveau des états transitoires X_2Y, pour lesquels une modélisation à un seul état n'est probablement pas suffisante étant donnée la dynamique de coarticulation. Il serait également utile d'enrichir les modèles syllabiques par l'information de durée moyenne des syllabes et par l'adaptation au genre du locuteur [9], ainsi que d'étendre le modèle de langage à plusieurs millions de mots. Finalement, une analyse plus poussée des stratégies d'élagage et des algorithmes de décodage permettrait de réduire la complexité de notre système.

BIBLIOGRAPHIE

[1] A. Acero, X. Huang. Augmented cepstral normalization for robust speech recognition. *Proc. of IEEE Automatic Speech Recognition Workshop*, 1995.

[2] L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Amer. Math. Soc.,* 73:360--363, 1967.

[3] J.M. Boite et C. Ris. Development of a french speech recognizer using a hybrid HMM/MLP system. *ESANN*, 1999.

[4] B. Fisher. Syllabification Software. Tsylb2-1.1. *http://www.nist.gov/speech/tools/*

[5] A. Ganapathiraju et al. Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE transactions on speech and audio processing*, vol. 9, N°4, May 2001.

[6] J.L. Gauvain, L.F. Lamel. Speaker-Independent Phone Recognition Using BREF. *ICASSP*, 1992.

[7] S. Greenberg. On the origins of speech intelligibility in the real world. *ESCARSR*, 97.

[8] L.F. Lamel, J.L. Gauvain, M. Eskénazi. BREF, a Large Vocabulary Spoken Corpus for French. *EUROSPEECH,* 1991.

[9] L.F. Lamel, J.L. Gauvain. High Performance Speaker-Independent Phone Recognition Using CDHMM. *EUROSPEECH*, 1993.

[10] B. Lowerre. The Harpy Speech Recognition System. *PhD Thesis, Carnegie-Mellon University,* 1976

[11] C. Pallier. Syllabation des représentations phonétiques de Brulex et de Lexique. 2004.

[12] L.R. Rabiner, B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January 4-15, 1986.

[13] C. Schrumpf et al. Syllable-based language models in speech recognition for English spoken document retrieval. *Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib,* 2005.

[14] S. Young et al. Token Passing : a Simple Conceptual Model for connected Speech Recognition Systems. *Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dept,* 1989.

[15] S. Young. The general use of tying in phoneme-based hmm speech recognisers. *IEEE*, 1992.

[16] S. Young et al. The HTK Book (for HTK version 3.3). *Cambridge University Engineering Department*, April 2005.

Reconnaissance de parole non native fondée sur l'utilisation de confusion phonétique et de contraintes graphèmiques

Ghazi Bouselmi, Dominique Fohr, Irina Illina, Jean-Paul Haton

Projet Parole, LORIA-CNRS & INRIA, "http ://parole.loria.fr/", BP 239, 54600 Vandoeuvre-lès-Nancy, France
{ bousselm,fohr,illina,jph }@loria.fr

ABSTRACT

This paper presents a fully automated approach for the recognition of non native speech based on acoustic model modification. For a native language (LM) and a spoken language (LP), pronunciation variants of the phones of LP are automatically extracted from an existing non native database. These variants are stored in a confusion matrix between phones of LP and sequences of phones of LM. This confusion concept deals with the problem of non existence of match between some LM and LP phones. The confusion matrix is then used to modify the acoustic models (HMMs) of LP phones by integrating corresponding LM phone models as alternative HMM paths. We introduce graphemic contraints in the confusion extraction process. We claim that prononciation errors may depend on the graphemes related to each phone. The modified ASR system achieved a significant improvement varying between 20.3% and 43.2% (relative) in "*sentence error rate*" and between 26.6% and 50.0% (relative) in "*word error rate*". The introduction of graphemic contraints in the phonetic confusion allowed improvements while using the word-loop grammar.

1. INTRODUCTION

La dégradation des performances des systèmes de reconnaissance automatique de la parole (SRAP) confrontés à un locuteur non natif est un problème bien connu. Une solution pour améliorer les performances des RAPs en présence de parole non native consiste à augmenter leur tolérance face à des variations de prononciation. Le problème consiste à introduire dans ces systèmes des connaissances relatives à ces variantes de pronciations. Différentes approches ont été envisagées et diffèrent selon la manière d'extraire ou d'intégrer ces connaissances dans le système de reconnaissance. Quelques unes de ces approches sont brièvement décrites ci-dessous.

Dans [6], des experts en phonétique ont établi des règles de réécriture phonétique pour quelques couples de langue parlée (LP) et langue maternelle (LM). Ces règles transforment les phonèmes de la LP en phonèmes de la LM. Ainsi, toutes le prononciations alternatives exprimées en termes de phonèmes de la LM sont ajoutées au lexique.

Dans [3], pour chaque phrase du corpus, un alignement forcé de la prononciation canonique et une reconnaissance phonétique (en termes de phonèmes de la LP) sont effectués. Ces deux transcriptions sont ensuite comparés afin d'extraire une confusion phonétique. Enfin, cette dernière sert à ajouter les prononciations alternatives de chaque mot dynamiquement durant la phase de reconnaissance.

Dans [4], une confusion phonétique est extraite d'une façon similaire à la précédente. Toutefois, la prononciation canonique est alignée avec une prononciation phonétique exprimée en termes de phonèmes natifs. Les deux SRAPs de la langue parlée et maternelle sont utilisés à cet effet. Par la suite, les modèles gaussiens des phonèmes natifs sont fusionnés avec ceux des phonèmes natifs avec lesquels ils ont été confondus, et ce pour chaque état des modèles de Markov sous-jacents (HMM).

Suite à une étude de prononciations non native réalisée avec des locuteurs français, nous avons mis en exergue deux principaux problèmes. Nous avons constaté que les locuteurs non natifs ont tendance à prononcer certains phonèmes de la LP comme des phonèmes de leur langue maternelle. Par exemple, certains phonèmes de la LP qui n'existent pas dans la LM sont souvent réalisés comme des phonèmes acoustiquement proches de la LM. C'est le cas de la consonne anglaise '[ð]' (dans le mot *the*) qui est souvent prononcée '[z]' par les locuteurs français. Le second problème, que nous avons constaté, est que la graphie d'un mot influence le locuteur non natif dans la façon de prononcer ce mot. Face à des mots qu'ils ne connaissent pas, ou encore des mots dont la graphie est identique dans les deux langues, ces locuteurs ont souvent tendance à réaliser une prononciation similaire à leur LM. Pour illustrer cela, considérons l'exemple de la table 1 où sont présentées les prononciations canoniques et les prononciations réalisées par des locuteurs français pour les mots anglais "*approach*" et "*position*". La table 1 montre que le phonème anglais '[ə]' est réalisé comme le phonème français '[a]' lorsqu'il correspond au caractère 'a' et comme le phonème français '[ɔ]' lorsqu'il correspond au caractère 'o'.

TAB. 1: *Prononciation du phonème anglais ə*

Mot	Prononciation canonique (phonèmes anglais)	Réalisation acoustique par des français (phonèmes français)
Approach	**[ə]** [p] [r] [əʊ] [tʃ]	**[a]** [p] [r] [ɔ] [t] [ʃ]
Position	[p] **[ə]** [z] [i] [ʃ] **[ə]** [n]	[p] **[ɔ]** [z] [i] [ʃ] **[ɔ]** [n]

Une des difficultés à laquelle sont confrontés les locuteurs non natifs est que certains phonèmes de la LP n'ont pas de correspondant directs dans la LM. C'est le cas de la diphtongue anglaise '[aɪ]' qui peut être réalisée comme la suite de phonèmes italiens '[a] [i]'. Dans notre approche [1], nous avons introduit un nouveau concept de confusion

phonétique qui associe à un phonème de la LP une suite de phonèmes de la LM. Ce concept sera brièvement décrit dans les sections suivantes.

Dans cet article, nous introduisons la contrainte graphémique à la confusion phonétique. Nous supposons que prendre en compte la graphie dans la confusion phonétique peut améliorer les performances de la reconnaissance.

2. NOUVELLE APPROCHE

Nous rappelons brièvement notre méthode déjà décrite dans [1]. La confusion phonétique considérée met en jeu des phonèmes de la langue parlée et maternelle. En effet, les locuteurs non natifs tendent à réaliser les phonèmes comme dans leur langue maternelle. Nous avons donc introduit un nouveau concept de confusion qui associe une suite de phonèmes de la langue maternelle aux phonèmes prononcés.

Pour chaque phrase du corpus d'adaptation, nous effectuons :
- un alignement forcé du signal audio avec la suite de modèles acoustiques de la langue parlée correspondant à la transcription canonique de la phrase,
- une reconnaissance phonétique du signal audio avec des modèles acoustiques de la langue maternelle.

Nous comparons ces deux transcriptions afin d'extraire des règles de confusion phonétique. Ces règles sont ensuite utilisées pour modifier les modèles de Markov (HMM) des phonèmes du SRAP de la langue parlée. Les HMMs correspondant à la séquence de phonèmes natifs sont ajoutés comme chemins alternatifs dans le HMM du phonème de la langue parlée. Nous supposons que cette utilisation de la confusion minimise le surcoût de puissance de calcul pour le nouveau SRAP. En effet, la modification du lexique (ajout de toutes les prononciations possibles) peut induire un surcoût très important ([2]). De plus, la modification des modèles gaussiens, comme dans [5], peut nuire à la cohérence temporelle des modèles acoustiques.

2.1. Extraction des règles de confusion

Deux ensembles de modèles acoustiques, ceux de la LP et de la LM, sont utilisés pour extraire des règles de confusion phonétique. Pour chaque phrase prononcée par un locuteur non natif, nous effectuons :
- un alignement phonétique forcé avec les phonèmes de la LP (prononciation canonique)
- une reconnaissance phonétique avec des phonèmes de la LM. Une simple boucle de phonèmes est utilisée à cet effet.

La comparaison des deux transcriptions permet ensuite de déduire les associations entre les phonèmes de la LP et les séquences de phonèmes de la LM. Un phonème $[K]$ de la langue parlée est associé à la suite de phonèmes $(M_i)_{i\in I}$ si chaque phonème M_i est inclus (pour plus de la moitié) pendant la durée de la prononciation du phonème $[K]$.

La prochaine étape consiste à extraire les règles de confusion à partir de ces associations. Seules les règles les plus pertinentes sont retenues. L'estimation au maximum de vraisemblance de la probabilité des règles ($P(K ==> (M_i)_{i\in I})$) est calculée pour chaque phonème $[K]$ (de la LP). Seules les règles dont la probabilité est supérieure à un seuil arbitraire α sont retenues.

Voici un exemple de règles données par notre système pour la diphtongue anglaise [aɪ] (où la langue maternelle est l'italien) :

[aɪ]==>[a] [i] $P([aɪ]==>[a]\ [i]) = 0.6$
[aɪ]==>[a] [e] $P([aɪ]==>[a]\ [e]) = 0.4$

Les mêmes règles ont été extraites dans le cas où la langue maternelle est l'espagnol et le grec.

2.2. Modification des modèles acoustiques

Les modèles HMMs du SRAP de la LP sont modifiés à l'aide des règles de confusion extraites à l'étape précédente. Pour chaque phone $[K]$ de la langue parlée, un chemin alternatif est ajouté au modèle HMM de $[K]$ (SRAP de la LP). Pour chaque règle $r \in R_K$ (règles sélectionnées pour le phonème $[K]$), un chemin correspondant à la partie droite de la règle est rajouté au modèle HMM de $[K]$. Ce nouveau chemin est la concaténation des modèles HMM (SRAP de la LM) correspondant aux phonèmes de la partie droite de r.

Nous utilisons une pondération entre le modèle acoutique de la LP et ceux de la LM. Ici β correspond au poids du modèle original. Dans le modèle HMM modifié, la transition liant l'état non émetteur de départ au modèle original a une probabilité β. De même, la probabilité liant cet état non émetteur à chaque chemin ajouté (pour une règle $r \in R_K$) est calculée comme suit :

$$P'(r) = (1-\beta)\frac{P(r)}{\Sigma_{x\in R_K}(P(x))} \quad (1)$$

Étant données les règles de confusion décrite dans le paragraphe 2.1 pour la diphtongue anglaise [aɪ], on obtient le modèle HMM représenté dans la figure 1.

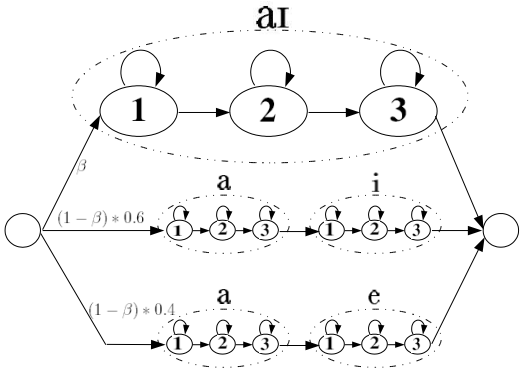


FIG. 1: *HMM modifié pour le phonème anglais [aɪ].*

3. CONTRAINTES GRAPHEMIQUES

Comme nous l'avons expliqué précédemment, les erreurs réalisées par les locuteurs non natifs sont fortement liées à la graphie des mots. Nous supposons donc que la prise en compte des contraintes graphémiques dans la confusion phonétique est susceptible d'apporter une amélioration de performances des SRAP.

La finalité de cette étape est d'associer automatiquement

les phonèmes à leur graphie sous-jacente pour chaque mots du dictionnaire.

3.1. Alignement automatique phonème-graphème

Étant données la graphie et la prononciation canonique d'un mot, le but est d'associer les phonèmes aux graphèmes correspondants. Cependant, l'information recherchée est l'alignement phonème-graphème et non pas une traduction graphème-phonème.

Un système HMM discret a été utilisé pour effectuer cette tâche. Ce système discret a été entraîné sur le dictionnaire phonétique du CMU. Les observations discrètes représentent les graphèmes et les modèles HMMs (mono-état) représentent les phonèmes. Le système HMM discret peut être utilisé pour effectuer l'alignement phonème-graphème sur le dictionnaire du SRAP cible, via l'algorithme de Viterbi.

Mise en place du système HMM discret
L'alignement phonème-graphème est extrait d'une manière automatique à partir d'un grand dictionnaire phonétique. Dans notre système HMM discret, les observations discrètes représentent les graphèmes, les modèles HMM représentent les phonèmes, et le dictionnaire phonétique représente le corpus d'apprentissage.

Les modèles HMM discrets initiaux ont une probabilité d'émission uniforme pour tous les symboles discrets (correspondant chacun à un graphème). Enfin, pour chaque mot du dictionnaire d'apprentissage, un fichier de données discrètes correspondant à la suite de symboles discrets (graphèmes du mot) est créé.

Donc, un alignement *phonétique* est effectué sur le dictionnaire d'apprentissage afin de déterminer les associations entre les phonèmes et les graphèmes. Seules les associations les plus représentées sont retenues. Ceci évite les erreurs dues à un alignement erroné ou encore à une inconsistance dans le dictionnaire d'apprentissage lui même. Une association a_K relative à un phonème $[K]$ n'est retenue que si elle satisfait l'inéquation 2.

$$N(a_K) \geq \gamma \, \Sigma_{a'_K \in A_K} N(a'_K) \qquad (2)$$

où A_K est l'ensemble des associations phonème-graphème pour le phonème $[K]$, $N(a_K)$ le nombre d'apparition de a_K, et γ une constante arbitraire.

Utilisation des contraintes graphémiques
Nous avons adopté une méthode simple pour appliquer les contraintes graphèmiques au SRAP cible. Nous proposons de remplacer les phonèmes dans la prononciation des mots du lexique par le couple (phonème,graphème) pour chacun des mots du dictionnaire (SRAP cible). La prononciation d'un mot n'est plus une suite de phonème, mais une suite de couple (phonème,graphèmes). L'exemple suivant illustre ceci pour le mot anglais "*speech*" :

ancienne prononciation	[s] [p] [i :] [tʃ]
nouvelle prononciation	[s]-S [p]-P [i :]-EE [tʃ]-CH

A cet effet, un alignement forcé sur le dictionnaire du SRAP cible est effectué en utilisant le système HMM discret entraîné (voir section précédente). Ceci fourni les associations phonème-graphème pour chaque mot du dictionnaire cible. Si une association phonème-graphème n'apparaît pas dans la liste des associations retenues (voir section précédente), le phonème reste sans contrainte graphémique dans la prononciation du mot.

La dernière modification consiste à ajouter les modèles HMM correspondant aux nouveaux phonèmes considérés. Pour chaque phonème $[K]$ avec une contrainte graphèmique X, $[K]$ est cloné en un nouveau phonème $[K]$-X.

3.2. Problèmes d'alignement

Dans un mot, un graphème peut être lié à plusieurs phonèmes. C'est le cas du mot anglais "*used*", prononcé [i] [u] [z] [d]. L'application directe de la méthode décrite plus haut donnerait l'alignement unique suivant : [i]-U, [u]-S, [z]-E et [d]-D, qui est bien évidemment faux. En effet, les observations ne peuvent pas être partagées entre les états dans un système HMM. De même, il n'est pas possible de considérer les phonèmes comme étant les observations puisqu'un phonème peut être liés à plusieurs graphèmes. Pour résoudre ce problème, nous avons opté pour la duplication des observations. Par exemple, pour le mot "*used*", la suite de symboles (U, U, S, S, E, E, D, D) sera considérée. Nous obtenons ainsi l'alignement ([i]-U, [u]-U, [z]-SS, [d]-EEDD). Un traitement postérieur donnera l'alignement correct : ([i]-U, [u]-U, [z]-S, [d]-ED).

4. EXPÉRIMENTATIONS

Notre travail a été effectué dans le cadre du projet Européen *HIWIRE* dont le but est d'améliorer les performances des SRAP dans les environnements mobiles et bruités. Le projet *HIWIRE* vise le développement d'un système automatique pour le contrôle vocal d'avions par les pilotes.

4.1. Conditions expérimentales

Notre corpus se compose de 4 parties : française, italienne, grèque et espagnole comportant respectivement 31, 20, 20 et 10 locuteurs. Chacun de ces locuteurs a prononcé 100 phrases en anglais. Nous avons considéré l'union de ces 4 parties comme un corpus international sur lequel nous avons effectué des tests de reconnaissance. La parole a été enregistrée à une fréquence de 16KHz. La paramétrisation MFCC utilisée consiste en 13 paramètres statiques avec leur dérivées premières et secondes. Les 46 monophones anglais ont été entraînés sur la base de donnés *TIMIT*. Les 40 monophones français ont été entraînés sur le corpus *ESTER*, composées de 90 heures de bulletins d'informations radiophoniques. Les modèles acoustiques espagnols, grecs et italiens ont été entraînés sur des corpus de parole espagnol, grec et italien (respectivement). Les modèles gaussiens des HMM possèdent 128 gaussiennes et des matrices de covariances diagonales. Le vocabulaire comporte 134 mots et la grammaire est un langage de commande. Une grammaire libre a également été utilisée (boucle de mots). Nous avons utilisé les 50 premières phrases de chaque locuteur pour l'extraction de la confusion, et les 50 dernières pour les tests.

4.2. Tests et résultats

Nous avons testé le système de référence (sans aucune adaptation) et le système de "confusion phonétique" avec

TAB. 2: *Résultats des tests effectués sur corpus français, italien, espagnol et grec (en %).*

	français		italien		espagnol		grec	
Système	WER	SER	WER	SER	WER	SER	WER	SER
grammaire contrainte :								
- système de référence	6.0	12.8	10.5	19.6	7.0	14.9	5.8	13.2
- "confusion phonétique"	**4.4**	**10.2**	**6.9**	**14.1**	**5.1**	**11.8**	**2.9**	**7.5**
- "confusion phonétique" + contraintes graphèmiques	4.9	11.3	8.2	15.9	6.2	13.6	6.3	15.8
grammaire libre :								
- système de référence	37.7	47.9	45.5	52.0	39.9	53.5	36.7	49.2
- "confusion phonétique"	27.3	42.1	31.3	46.2	**31.3**	**44.5**	**20.2**	**34.9**
- "confusion phonétique" + contraintes graphèmiques	**26.2**	**41.9**	**30.5**	**45.5**	**31.3**	46.5	57.0	76.2

la grammaire contrainte et la grammaire libre. Nous avons effectué des tests séparés sur les corpus français, italien, espagnol, grec et international. Nous avons extrait des règles de confusion entre les phonèmes anglais et les phonèmes de la langue native respective de chaque partie de notre corpus. Toutefois, pour les tests avec le corpus international, nous avons extrait des confusions phonétiques uniquement avec des phonèmes de la langue parlée : confusion entre prononciation canonique (anglais) et phonèmes réellement prononcés (anglais).

Comme le montrent les tableaux 2 et 3, la confusion phonétique apporte une amélioration des taux de reconnaissance sur tous les corpus. Pour les tests utilisant la confusion phonétique et la grammaire contrainte, cette amélioration varie de 26.6% à 50.0% (relatif) en WER (taux d'erreurs en mots) et de 20.3% à 43.2% (relatif) en SER (taux d'erreurs en phrases). Les tests sur la confusion phonétique avec la grammaire libre affichent des améliorations variant de 21.6% à 45.0% en WER et de 11.2% à 29.1% en SER. En revanche, l'ajout des contraintes graphémiques à la confusion phonétique n'a pas eu de répercussions positives sur les taux de reconnaissances (par rapport à la confusion phonétique seule) en ce qui concerne les tests effectués avec la grammaire contrainte. Néanmoins, nous observons une légère amélioration pour les tests impliquant la grammaire libre (par rapport à la confusion phonétique seule) tant en WER qu'en SER. Nous pensons que cette dégradation est dûe à ce que la grammaire contrainte guide très bien la reconnaissance et donc l'ajout des contraintes graphèmiques n'amèliore pas les taux. En effet, il s'agit d'un langage de commande strict composé de seulement 134 mots.

TAB. 3: *Résultats des tests effectués sur le international (en %).*

Système	WER	SER
grammaire contrainte :		
- système de référence	7.1	14.5
- "confusion phonétique"	**5.7**	**12.1**
- "confusion phonétique" + contraintes graphèmiques	5.8	12.4
grammaire libre :		
- système de référence	38.5	49.9
- "confusion phonétique"	31.2	44.8
- "confusion phonétique" + contraintes graphèmiques	**30.2**	**43.7**

5. CONCLUSION

Nous avons présenté une nouvelle approche pour l'amélioration de la reconnaissance automatique de la parole prononcée par des locuteurs non natifs. Elle est basée sur l'utilisation de la confusion phonétique et des contraintes graphèmiques. Dans notre approche, les phonèmes de la langue parlée sont associés à une suite de phonèmes de la langue maternelle du locuteur. Nous avons présenté une nouvelle utilisation de la confusion phonétique qui consiste à ajouter de nouveaux chemins alternatifs dans les modèles HMM des phonèmes de la langue parlée. Nous avons également proposé l'adjonction des contraintes graphèmiques à la confusion phonètique. En effet, les erreurs de prononciations des phonèmes sont fortement corrélées à la graphie des mots. La confusion phonétique a apporté des améliorations significatives dans les taux de reconnaissance sur notre corpus. Toutefois, l'ajout des contraintes graphémiques n'a été favorable que lors de l'utilisation d'une grammaire non contrainte.

6. REMERCIEMENTS

Ce travail a été partiellement financé par le projet Européen *HIWIRE* (Human Input that Works In Real Environments), contrat numéro 507943, "sixth framework programme, information society technologies".

RÉFÉRENCES

[1] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton. Fully automated non-native speech recognition using confusion-based acoustic model integration. In *In Proc. Eurospeech/Interspeech*, 2005.

[2] S. Goronzy, R. Kompe, and S. Rapp. Generating non-native pronunciation variants for lexicon adaptation. In *Eurospeech*, 2001.

[3] K. Livescu and J. Glass. Lexical modeling of non-native speech for automatic speech recognition. In *ICASSP*, 2000.

[4] J. Morgan. Making a speech recognizer tolerate non-native speech through gaussian mixture merging. In *InSTIL/ICALL*, 2004.

[5] P. Nguyen, P. Gelin, J.-C. Junqua, and J.-T. Chien. N-best based supervised and unsupervised adaptation for native and non-native speakers in cars. In *ICASSP*, 1999.

[6] Stefan Schaden. Generating non-native pronunciation lexicons by phonological rule. In *ICSLP*, 2004.

Etude par transillumination des consonnes occlusives simples et géminées de l'arabe marocain

Chakir Zeroual[1&2], Phil Hoole[3] et Susanne Fuchs[4]

1. Université Sidi Mohamed Ben-Abdellah, Faculté Polydisciplinaire de Taza, BP. 1223 Taza, Maroc.
2. Laboratoire de Phonétique et Phonologie, Sorbonne-Nouvelle/CNRS-UMR7018, Paris-France.
3. Institut fur Phonetik, Munich, Germany.
4. ZAS/Phonetik. Jaegerstr. 10-11. 10117, Berlin, Germany.

Chakirzeroual@yahoo.fr ; hoole@phonetik.uni-muenchen.de. ; fuchs@zas.gwz-berlin.de

ABSTRACT

This study provides an acoustical and transillumination analysis of the laryngeal gestures responsible for VOT differences between voiceless plosives in Moroccan Arabic. The abduction and adduction phases and the interval between the maximal glottal opening and the release (MGO-REL) are longer during the geminate than during simple plosives. MGO-REL is shorter during the aspirated plosives than during unaspirated ones. Geminate plosives have a closure duration, a total duration and MGO that are larger than their simple correspondents. The closure duration is longer during unaspirated plosives than during aspirated ones. The voiceless geminate plosives have the same values of the VOT as their simple counterparts.

1. INTRODUCTION

L'arabe marocain (AM) possède huit consonnes occlusives (Tableau 1) qui peuvent être simples ou géminées. Une analyse précédente [15] assez exhaustive (données de 10 locuteurs) a montré qu'à l'initiale de mot les occlusives simples [b d g D] ont un VOT négatif et [t T k q] un VOT positif dont la durée varie de manière assez importante ([k] : 59msec, [t] : 61msec, [q] : 37msec, [T] : 22msec).

Tableau 1 : Consonnes occlusives de l'AM. Emph = Emphatique et renvoie à l'articulation secondaire "d'uvularisation" qui s'ajoute à [t d] pour donner [T D].

	Bilabiale	Coronale	Vélaire	Uvulaire
non-Emph.	b	t d	k g	q
Emph.		T D		

[t k T q] de l'AM sont assez particulières. En effet, même si [t k] possèdent un VOT très long elles n'ont pas les mêmes propriétés auditives des occlusives aspirées de l'anglais). Cette caractéristique existe aussi dans plusieurs dialectes arabes, et la caractérisation phonétique précise de [t k T q] n'est pas la même chez tous les auteurs [15, 16].

Les analyses acoustiques ne sont pas suffisantes pour l'identification des trois événements acoustiques qui peuvent être observés durant le bruit de relâchement (explosion + friction + aspiration) des occlusives sourdes. En effet, la séparation entre ces événements sur les spectrogrammes n'est pas très évidente. D'autres analyses phonétiques sont donc utiles pour mieux caractériser phonétiquement ces occlusives sourdes. Dans la suite de cette présentation, nous considérons [t k] comme des occlusives aspirées et [T q] des occlusives non aspirées. "Aspiré" est, ici, synonyme de "bruit de relâchement (ou VOT positif) très long". Cette interprétation est basée sur les durées du VOT (tableau 2) de [t k q T] produites par le locuteur qui a participé aux expériences phonétiques discutées dans ce travail.

Le but principal de cette étude est de caractériser les ajustements articulatoires laryngaux des occlusives sourdes simples et géminées. Nous essayerons d'expliquer leurs relations avec les ajustements supralaryngaux pour rendre comptes des différences observées au niveau du VOT durant ces consonnes. Les occlusives voisées seront analysées dans un travail séparé. Cette étude se base aussi sur les résultats des analyses acoustiques et physiologiques précédentes [14, 15]. Elle se démarque de ces dernières par deux aspects principaux : (i) utilisation de la transillumination qui permet une identification plus précise des différentes phases du geste glottique ; (ii) comparaison entre les propriétés acoustiques et physiologiques des occlusives simples et géminées.

2. MATERIEL ET METHODE

Durant une expérience par transillumination, un endoscope flexible a été inséré à travers les fosses nasales d'un locuteur (38 ans). Sur la surface externe de son cou ont été placées deux cellules photoélectriques PGG1 (entre les cartilages thyroïde et cricoïde) et PGG2 (en bas du cricoïde) pour capter la quantité de lumière, émise par l'endoscope, qui passe à travers la glotte et qui est proportionnelle à son degré d'ouverture. L'analyse a été faite sur les signaux obtenus par PGG2 puisqu'ils étaient plus stables (moins influencés par les mouvements du larynx et de l'épiglotte). Durant cette expérience, des enregistrements vidéo ont été également effectués.

Ce locuteur a répété plusieurs fois un corpus composé de mots et de quelques non-mots de l'AM contenant presque toutes les consonnes simples (contexte [-iCi]) et géminées (contexte [-ɪCCi-]). Nous présentons ici les résultats de [t tt T TT k kk q qq].

Les films vidéo ont été analysés par Adobe Premiere7, les images par Adobe Photoshop7, les données audio par Praat, les données transillumination par Matlab et les analyses statistiques par StatView.

3. RESULTATS ET DISCUSSION

Bien que [t k q T] simples et géminées soient sourdes, les tracés par transillumination montrent que seules [t k q] développent, à l'intervocalique, des phases d'abduction et d'adduction clairement définies. La fibroscopie montre que [t k q] simples et géminées sont produites avec une ouverture glottique assez importante, accompagnée d'un écartement entre les aryténoïdes. Par contre, durant [T TT], les aryténoïdes restent collés, et seule la partie antérieure de la glotte s'ouvre très légèrement. Il semble que seules [t tt k kk q qq] possèdent un geste actif d'abduction. L'abduction de la partie antérieure de la glotte durant [T TT] serait passive provoquée très probablement par l'augmentation de la pression intraorale (Po). Ces hypothèses sont développées dans un travail séparé comprenant aussi l'analyse acoustique et physiologique des occlusives voisées simples et géminées.

Nous présentons, ici, une analyse acoustique assez exhaustive de [t tt k kk q qq T TT] (Tableau 2) qui regroupent (i) la mesure de la durée de la phase d'occlusion (OCC), (ii) la durée du VOT (VOT) et (iii) la durée totale (DTL : occlusion+VOT). Par contre l'analyse du cycle glottique (abduction+adduction) concernera uniquement [t tt k kk q qq]. Elle comporte (iv) la mesure de la durée de l'intervalle entre le début de l'occlusion et le moment de l'ouverture glottique maximale (OCC-OGM), que nous considérons ici comme représentant la phase d'abduction, (v) la durée de l'intervalle entre l'OGM et le relâchement (OGM-REL), (vi) ainsi que la durée entre l'OGM et le début de la voyelle suivante (OGM-VOY), que nous considérons ici comme représentant la phase d'adduction. Nous avons également mesuré (unités arbitraires) (vii) le degré de l'ouverture glottique maximale (OGM), ainsi que (viii) la vélocité du geste d'adduction (VEL).

3.1 Consonnes simples

Trois tests séparés ANOVA à un seul facteur montrent que la durée de l'occlusion (OCC : [F(7, 32) = 116,77 ; p<0,001]), la durée totale (DTL : [F(7, 32) = 90,13 ; p<0,001]) et celle du VOT ([F(7, 32) = 50,82 ; p<0,001]) varient de manière significative en fonction de la nature des consonnes [t tt k kk q qq T TT]. Dans cette présentation, nous nous baserons sur les Tests PLSD de Fisher pour interpréter les analyses a posteriori.

[t k] (p=0,29) et [T q] (p=0,12), comparées entre elles, montrent qu'elles ont des VOT statistiquement identiques. Cependant le VOT est beaucoup plus long durant [t k] que durant [T q] (p<0,001). La durée de l'occlusion est statistiquement similaire durant [T q] (p=0,18) et [t k] (p=0,90), mais significativement plus longue durant [T q] comparées à [t k] (p<0,01). Ce résultat rejoint les observations faites par d'autres auteurs [1 4 7] montrant que les occlusives sourdes non aspirées ont généralement une occlusion plus longue que les occlusives aspirées.

La durée totale (DTL) est statistiquement similaire durant [t k] (p=0,42) et [q T] (p=0,91). Elle est significativement supérieure durant [t k] comparées [q T] (p<0,025), même si ces dernières ont une occlusion plus longue que les premières. Cette différence est à attribuer à la durée très importante du VOT durant [t k]. Des différences par rapport à la durée de l'occlusion ont été mentionnées par d'autres auteurs [1 4]. Selon Maddieson [8] "the consonant gesture is timed in some way that directly relates to the time of the pressure peak, then broadly speaking, the further back in the oral cavity a stop closure is formed, the shorter its acoustic closure duration will be". Nos données vont à l'encontre de cette hypothèse qui prédit que l'occlusion de [q] serait la plus brève. Notons que des analyses aérodynamiques [14] montrent qu'en AM Po durant [t k q] n'est pas significativement différente.

Cinq autres tests séparés ANOVA à un seul facteur montrent que les cinq mesures relevées sur la courbe de la vélocité varient significativement en fonction des consonnes [t tt k kk q qq T TT] (OCC-OGM : [F(5, 24) = 42,76 ; p<0,001] ; OGM-REL : [F(5, 24) = 30,68 ; p<0.001]; OGM-VOY : [F(5, 24) = 34,86 ; p<0,001] et VEL : [F(5, 24) = 36,94 ; p<0,001]) et sur la courbe de l'aperture glottique (OGM : [F(5, 24) = 8,87 ; p<0,001]). Là aussi, nous nous baserons sur les tests PLSD de Fisher pour interpréter les analyses à posteriori.

La durée entre le début de l'occlusion et le moment de l'ouverture glottique maximal, que nous assimilons à la phase d'abduction, est statistiquement identique durant [t q k]. l'OGM est atteinte très légèrement avant le relâchement durant [t k] (p=0,61) et significativement plus en avant durant [q] (p<0,01). L'OGM est similaire durant [t k] (p=0,09), mais significativement plus grande durant [q] (p<0,001), malgré cela, le geste d'adduction durant [q] a une durée totale qui est significativement plus courte que [k] (p<0,05) et [t] (p<0,001). Deux hypothèses peuvent être avancées concernant les différences par rapport à la durée de l'adduction.

En effet, le degré de l'OGM et la vélocité du geste glottique peuvent être directement contrôlés en fonction des valeurs du VOT (voir aussi [5] et [7]). Cette hypothèse peut expliquer le comportement particulier de [q] qui possède une durée d'adduction plus courte malgré une OGM plus importante. Les valeurs de la vélocité présentent un argument en faveur de cette hypothèse, puisque [q] possède la valeur la plus importante de la vélocité comparée à [k] (p<0,05) et [t] (p<0,001).

L'autre explication pour l'adduction plus courte durant [q] est à rechercher au niveau des contraintes aérodynamiques pendant le relâchement. En effet, selon Stevens [13], plus la surface du contact supralaryngal est étendue plus le relâchement est retardé par la succion crée

par l'effet de Bernoulli**.** Ce dernier ralentit la baisse de Po qui elle-même ralentit la vitesse d'adduction des cordes vocales et allonge la durée du VOT. Les données aérodynamiques [14] montrent que Po baisse plus rapidement durant [q] comparée à [t k]. Ce résultat combiné avec l'hypothèse de Stevens suggèrent que [t k] seraient produites avec une surface de contact plus étendue, comparés à [q], ce qui rallonge indirectement la phase d'adduction de [t k]. Cet effet est plus marqué durant [t k], puisque l'adduction est amorcée immédiatement avant le relâchement. L'hypothèse d'une surface de contact moins étendue est proposée aussi par Cho et al [1] pour expliquer la durée très faible du VOT de [q] dans d'autres langues.

3.2 Consonnes géminées

La durée totale (DTL) et de la phase d'occlusion (OCC) de [tt kk qq TT] sont significativement plus grandes ($p<0,001$) que [t k q T] (rapports géminée/simple respectivement égal à 1,50 et 1,81). Nos analyses montrent que ce rapport géminée/simple, concernant DTL, est légèrement inférieur à celui obtenu, par exemple, pour des occlusives sourdes en arabe irakien (1,73) [3].

La durée du VOT reste statistiquement similaire lorsque nous passons de la forme simple à la forme géminée. Notons que le VOT des géminées est légèrement plus court comparées aux simples, cette baisse est plus importante durant [TT]. Des différences non significatives de la durée du VOT entre les occlusives simples et géminées, avec une valeur légèrement plus faible pour les géminées, apparaissent aussi dans les données d'autres travaux [2, 11]. Les différences au niveau du VOT entre les géminées sont parallèles à celles observées entre les simples : pas de différences entre [tt kk] ($p=0,17$) et entre [TT qq] ($p=0,31$), alors que le VOT est beaucoup plus long durant [tt kk] comparées à [T q] ($p<0,001$).

La différence, par rapport à la durée de l'occlusion, entre [TT] et [qq] ($p=0,124$) ainsi qu'entre [tt] et [kk] ($p=0,124$) reste non significative. Par contre, [TT qq] possèdent une durée de l'occlusion qui est significativement supérieure à celle de [tt kk] ($p<0,01$). Comme pour les occlusives sourdes simples, la durée de l'occlusion est donc plus longue durant les géminées non aspirées comparées aux géminées aspirées. Malgré ce dernier résultat, les différences entre les durées totales de [tt kk qq TT] ne sont pas statistiquement significatives.

L'OGM est atteinte beaucoup plus en retard durant [tt kk qq] que durant [t k q] ($p<0,001$), mais comme durant [t k q], l'OGM est atteinte pratiquement au même moment durant [tt kk qq]. Par rapport au relâchement l'OGM est atteinte en général beaucoup plus tôt durant les géminées que durant les simples ($p<0,001$), elle est aussi relativement plus proche du relâchement durant [tt kk] ($p=0,59$) comparées à [qq] ([tt kk] vs [qq], $p<0,02$).

Seules [tt] et [kk] présentent des OGM significativement plus importantes que leurs correspondantes simples [t] ($p<0,01$) et [k] ($p<0,01$), la différence entre [qq] et [q] reste non significative ($p=0,09$). La durée de la phase d'adduction ($p<0,001$) et la vélocité ($p<0,02$) sont plus importantes durant les géminées comparées à leurs correspondantes simples. Les valeurs plus importantes de la vélocité durant les géminées sont globalement attendues si nous considérons que globalement les géminées ont une OGM plus importante que les simples. En effet, les analyses articulatoires montrent, en générale, une corrélation positive entre l'amplitude et la vélocité du mouvement des gestes articulatoires [9].

Comme pour les consonnes simples, [qq] présente l'OGM la plus importante suivie de [kk] et enfin de [tt], toutefois seule la comparaison [qq] vs [tt] est significative ($p<0,04$). [kk] présente une vélocité qui est plus importante significativement de [qq] et non significativement de [tt]. Les tests à posteriori montrent que la durée de la phase d'adduction reste statistiquement similaire durant [tt kk qq], malgré les différences au niveau des valeurs de l'OGM et de la vélocité. Ce résultat nous rapproche de certaines observations faites au niveau des gestes supralaryngaux qui montrent "no systematic relationship between either displacement or maximum velocity and the duration of the gesture (Kent and Moll, 1969, 1972; Kent and Netsell, 1971; Kuehn and Moll, 1976; Ostry et al., 1983) [9].

4. CONCLUSION

Nos analyses acoustiques montrent que les occlusives géminées possèdent une durée de la phase d'occlusion et une durée totale qui sont très supérieures à celles de leurs correspondantes simples. Les occlusives sourdes non aspirées, simples ou géminées, ont une durée de l'occlusion qui est plus longue que leurs correspondantes aspirées. Cependant, les occlusives sourdes géminées ont les mêmes valeurs du VOT que leurs correspondantes simples.

Nos analyses articulatoires ont montré que les phases d'abduction et d'adduction et la durée entre l'ouverture glottique maximale et le relâchement sont plus longues durant les géminées comparées aux simples. Par contre, l'OGM est plus proche du relâchement durant les occlusives aspirées, simples ou géminées, comparées à leurs correspondantes non aspirées. L'OGM est généralement plus importante durant les occlusives sourdes géminées que durant leurs correspondantes simples.

L'analyse en cours d'autres données EMA devra nous aider à expliquer si les différences temporelles et spatiales observées au niveau du geste glottique sont dues aux ajustements laryngaux et/ou à l'action des contraintes aérodynamiques.

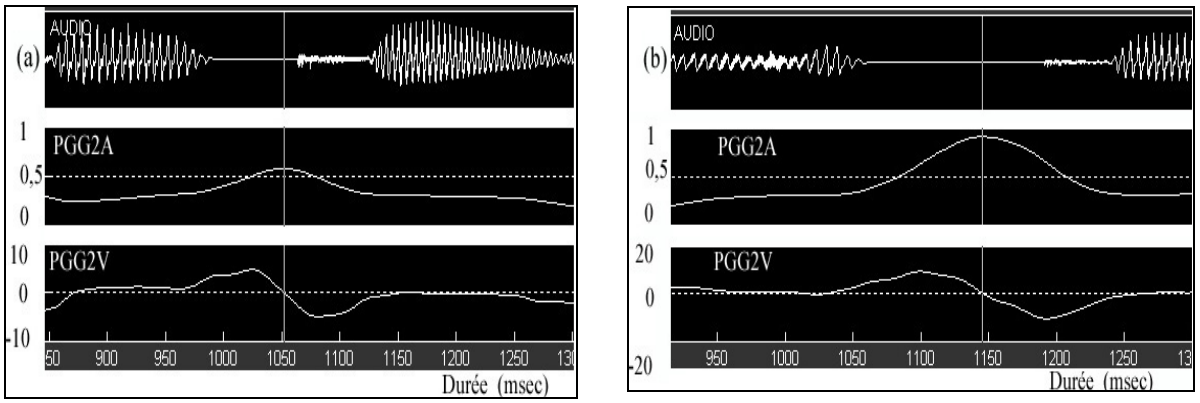


Figure 1 : Tracés de la courbe de l'aperture glottique (PGG2A) obtenue par transillumination, ainsi que la courbe de la vélocité (PGG2V) calculée sur la première durant les séquences [-iki-] (a) et [-ɨkki-] (b) extraites des items [fiki] et [zɨkkit] de l'AM. Notons que l'axe des 'y' de PGG2V de (a) et (b) n'a pas la même échelle.

Table 2 : Moyennes et écartypes des mesures articulatoires et acoustiques relevées durant [t k q T] simples dans [-iCi-] et géminées dans [-ɨCCi-] : durées (msec) des intervalles entre le début de l'occlusion et l'ouverture glottique maximale (OCC-OGM), entre l'OGM et le relâchement (OGM-REL), entre l'OGM et le début de la voyelle suivante (OGM-VOY), durée totale de la consonne (DTL), de l'occlusion (OCC), du VOT et enfin le degré de l'OGM et la vélocité de la phase d'adduction (VEL) [unités arbitraires]. Chaque valeur correspond à 5 répétitions produites par un seul locuteur.

	OCC-OGM	OGM-REL	OGM-VOY	OGM	VEL	DTL	OCC	VOT
[t]	73 (9)	-4 (9)	70 (6)	0,36 (0,05)	1,70 (0,58)	143 (5)	76 (6)	66 (4)
[k]	71 (4)	-7 (10)	68 (8)	0,50 (0,07)	3,44 (1,03)	138 (8)	77 (7)	61 (10)
[q]	71 (6)	-23 (13)	55 (10)	0,64 (0,09)	5,44 (1,59)	126 (6)	94 (11)	31 (10)
[T]				0,28 (0,04)		126 (7)	102 (5)	25 (2)
[tt]	107 (6)	-34 (6)	98 (8)	0,64 (0,09)	6,48 (1,87)	205 (11)	141 (7)	64 (5)
[kk]	105 (8)	-45 (5)	104 (8)	0,72 (0,19)	9,08 (2,34)	209 (14)	150 (9)	59 (5)
[qq]	105 (1,5)	-61 (9)	90 (2)	0,82 (0,16)	7,64 (0,38)	195 (3)	166 (8)	29 (7)
[TT]						191 (11)	174 (10)	17 (1)

BIBLIOGRAPHIE

[1] T. Cho, and P. Ladefoged. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27: 207-229, 1999.

[2] A. Cohn, W.H. Ham, and R.J. Podesva. The phonetic realization of singleton-geminate contrasts in three languages of Indonesia. *Proceedings of the XIVth ICPHS*, San Francisco, pp. 587-590, 1999.

[3] Z. M. Hassan. Gemination in Swedish and Arabic with a particular reference to the preceding vowel duration. An instrumental & comparative approach. *TMH-QPSR* Vol. 44: 81-84, 2002.

[4] P. Hoole, "Laryngeal Coarticulation. Section A: Coarticulatory investigations of the devoicing gesture". In: W.H Hardcastle and N. Hewlett (eds.), *Coarticulation: Theory, Data and Techniques.* Cambridge University Press, pp. 105-121, 1999.

[5] DW. Kim, H. Hirose, and S. Niimi. A fibroscopic study of laryngeal gestures for Korean intervocalic stops. Ann Bull RILP No.26: 13-29, 1992.

[6] P. Ladefoged and I. Maddieson. *Sounds of the worlds languages*. Oxford: Blackwells, 1996.

[7] A. Löfqvist. Interarticulator programming in stop production. *Journal of Phonetics* 8: 475-490, 1980.

[8] I. Maddieson. Phonetic Universals. In *The handbook of phonetic sciences* (J. Laver & W. J. Hardcastle, eds). Oxford: Blackwells, pp. 619-639, 1997.

[9] A. Parush, D.J. Ostry, and K.G. Munhall. A kinematic study of lingual coarticulation in VCV sequences. *J. Acoust. Soc. Am.* 74 (4) : 1115-1125, 1983.

[10] M. Pétursson. Aspiration et activité glottale. Examen expérimental à partir de consonnes islandaises. *Phonetica* 33 : 169-198, 1976.

[11] R. Ridouane. Geminates *vs.* Singleton Stops in Berber : An Acoustic, Fiberscopic and Photoglottographic study. *Proceedings of the XVth ICPHS*, Barcelona, pp. 1743-1746, 2004.

[12] M. Sawashima. and H. Hirose. Laryngeal gestures in speech production. *Ann. Bull. RILP*. 14: 29-51, 1980.

[13] K. N. Stevens. *Acoustic phonetics*. Cambridge: MIT Press, 1999.

[14] C. Zeroual. "Etude aérodynamique et acoustique des occlusives emphatiques et non-emphatiques de l'arabe marocain", *XXIVèmes Journées d'Etude sur la Parole*, 24-27 June, Nancy, France, pp. 365-368, 2002.

[15] C. Zeroual. *Propos controversés sur la phonétique et la phonologie de l'arabe marocain*. Thèse de Doctorat, Université de Paris 8, 2000.

[16] C. Zeroual, Affrication. *In:* M. Eid, A. Elgibali, K. Versteegh, M. Woidich, and A. Zaborski, ed. *Encyclopedia of Arabic Language and Linguistics*. Leiden: Brill Academic Publishers, 2005.

Analyse fibroscopique des consonnes sourdes en berbère

Rachid Ridouane

ENST-TSI/CNRS-LTCI UMR 5141
46, rue Barrault, 75634 Pris cedex 13
Tel : (33) 01 45 81 71 90, Fax : (33) 01 45 88 79 35
rachid.ridouane@wanadoo.fr

ABSTRACT

This article deals with laryngeal adjustments during the production of singleton voiceless consonants in Tashlhiyt Berber. It focuses on the influence of place and manner of articulation and effects of position in the word. Results provide evidence that the degree of glottal opening as well as the velocity of abduction-adduction gestures vary according to the place of articulation of stops and fricatives and their position in the word. Systematic differences, reflecting a universal tendency, were also observed between stops and fricatives. The specific laryngeal adjustments during the production of uvulars and so-called pharyngeal will be briefly outlined in the discussion.

1. INTRODUCTION

Pendant la parole, les cordes vocales sont parfois en abduction pour satisfaire les contraintes aérodynamiques nécessaires pour la production des occlusives ou des fricatives sourdes. Cette abduction est produite par l'écartement des cartilages aryténoïdiens auxquels sont fixées les extrémités postérieures des cordes vocales à l'arrière du larynx. Des ajustements spécifiques des cordes vocales sont produits selon la nature phonétique et phonologique de ces segments. Ces ajustements affectent entre autres l'amplitude de l'ouverture glottale et le rapport temporel entre les gestes glottaux et les gestes supraglottaux.

Cette étude traite de la nature de ces ajustements en berbère chleuh en s'intéressant plus particulièrement aux effets de la position (initiale, intervocalique, et finale) et du lieu d'articulation. Différents segments seront examinés (/t/, /tˤ/, /k/, /q/, pour les occlusives, et /f/, /s/, /ʃ/, /χ/, et /ħ/ pour les fricatives), avec une attention particulière sur ceux peu étudiés jusque là, et notamment la dentale emphatique, les uvulaires et la pharyngale. Il s'agira aussi, à travers ce travail, de fournir des points de comparaison entre nos résultats et les résultats obtenus à partir d'autres langues, comme l'anglais, le danois, l'islandais, le japonais et l'arabe marocain. L'objectif est de dégager, s'il y a lieu, des caractéristiques laryngales qui peuvent être considérées comme étant universelles.

2. METHODE & MATERIEL

La fibroscopie a été utilisée comme méthode d'investigation expérimentale. Un fibroscope de type Olympus ENF-P3 a été introduit par la narine et stabilisé à quelques millimètres de la glotte, ce qui a permis d'observer directement les mouvements des cordes vocales et des cartilages aryténoïdiens ainsi que certains mouvements de l'épiglotte. Une caméra Sony (XC-999 P) a été fixée sur le bout externe du fibroscope pour enregistrer un film vidéo sur magnétoscope "U-Matic" (VO-5800 PS). Un « micro-cravate » Sony a été utilisé pour l'enregistrement simultané du son, ce qui permet la synchronisation du son avec les images. L'acquisition du film vidéo (25 i/s) a été effectuée à l'aide d'un ordinateur PC équipé de la carte Miro DC 30 et du logiciel Adobe Première 5.1. L'analyse des données a été principalement faite en utilisant le logiciel SoundForge 5.0. qui permet d'avoir aussi bien le signal acoustique que les séquences vidéos. Les images, copiées à partir de ce logiciel, ont été traitées en utilisant Adobe Photoshop 5.0 et Adobe Illustrator 7.0.

Vu le caractère assez contraignant d'une prise de données par fibroscopie, et à l'image de beaucoup d'études ayant adopté cette technique (voir [1] pour une revue), un seul locuteur natif du berbère chleuh (l'auteur) a participé à cette expérience. Les données enregistrées sont des mots réels où les consonnes sourdes (/t/, /tˤ/, /k/, /q/, /f/, /s/, /ʃ/, /χ/, et /ħ/) apparaissent dans trois contextes différents : initiale, intervocalique et finale. Chaque segment cible a été placé à l'adjacence de la voyelle /i/ (#Ci, iCi, iC#) pour s'assurer que la racine de la langue ou l'épiglotte n'obstruent pas le fibroscope et empêcher une meilleure observation des mouvements laryngaux. Chaque forme a été répétée 5 fois dans 4 sessions différentes. Ces formes n'ont pas été mises dans une phrase cadre pour s'assurer que les segments cibles apparaissent bel et bien en positions initiale et finale absolues.

3. RESULTATS

Les résultats seront présentés en deux sections correspondant aux deux modes d'articulation examinés. Chaque section est subdivisée en trois sous-sections correspondant aux trois contextes analysés. L'intérêt sera porté principalement sur les différences qualitatives au niveau de la posture

globale du larynx et au niveau du cycle d'ouverture-fermeture de la glotte.

3.1. Les occlusives sourdes

La figure 1 indique la durée et le degré d'ouverture glottale pour chaque occlusive sourde dans les trois positions. Il s'agit de la moyenne des mesures effectuées sur 5 répétitions pour chaque forme lors d'une même session. Sachant que les mesures du degré de l'ouverture glottale obtenues par cette technique ne sont pas calibrées, les comparaisons de ce paramètre entre deux formes enregistrées pendant deux sessions différentes ne peuvent être qu'approximatives

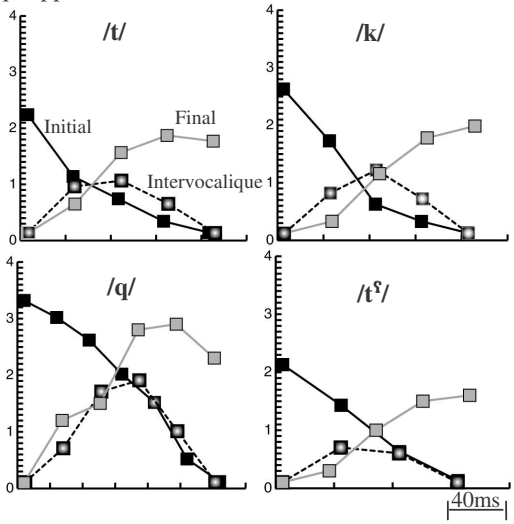


Figure 1. *Degré et durée d'ouverture glottale des occlusives dans les trois postions. La durée d'occlusion en position initiale a été déterminée en se basant sur les mesures du flux d'air oral obtenues à partir des mêmes formes (voir [1]).*

3.1.1. Position initiale

Seul le geste de fermeture glottale est visible dans cette position. Ce geste correspond à la transition entre « le mode respiratoire » et « le mode phonatoire ». Les occlusives dans cette position sont produites avec un degré d'ouverture glottale important qui se referme d'une manière plus ou moins progressive jusqu'à l'adduction totale au moment de la tenue de la voyelle qui suit. Quelques variations entre différents types d'occlusives ont été observées. L'uvulaire se réalise systématiquement avec un degré et une durée d'ouverture glottale plus importants. La dentale emphatique est produite, quant à elle, avec la plus faible ouverture glottale au moment du relâchement /t/ et /k/ semblent être réalisées avec les mêmes ajustements glottaux (le même degré d'ouverture glottale au moment du relâchement oral).

3.1.2 Position intervocalique

Les occlusives intervocaliques sont produites avec un geste balistique d'ouverture-fermeture de la glotte. /t/ et /k/ présentent pratiquement les mêmes configurations ; la glotte qui était fermée pendant la réalisation de la voyelle précédente s'ouvre progressivement pour atteindre son niveau maximal vers le relâchement oral. Elle entame ensuite, d'une manière progressive, sa phase fermante pour atteindre une fermeture complète à l'onset de la voyelle qui suit. Comme en position initiale, /q/ est systématiquement produit avec une durée et une amplitude d'ouverture glottale plus importantes. La glotte atteint son ouverture maximale loin du relâchement oral et se referme beaucoup plus rapidement. La dentale emphatique présente la plus faible amplitude glottale.

3.1.3 Position finale

Seul le geste d'ouverture glottale est visible dans cette position ; la glotte qui était fermée pendant la tenue de la voyelle précédente s'ouvre au début de l'occlusion et continue à s'ouvrir jusqu'à la phase respiratoire. Comme en position initiale et intervocalique, /q/ et /tˤ/ sont respectivement réalisés avec la plus large et la plus faible ouverture glottale.

3.2 Les fricatives sourdes.

La figure 2 indique la durée et le degré d'ouverture glottale pour chaque fricative sourde selon les trois positions.

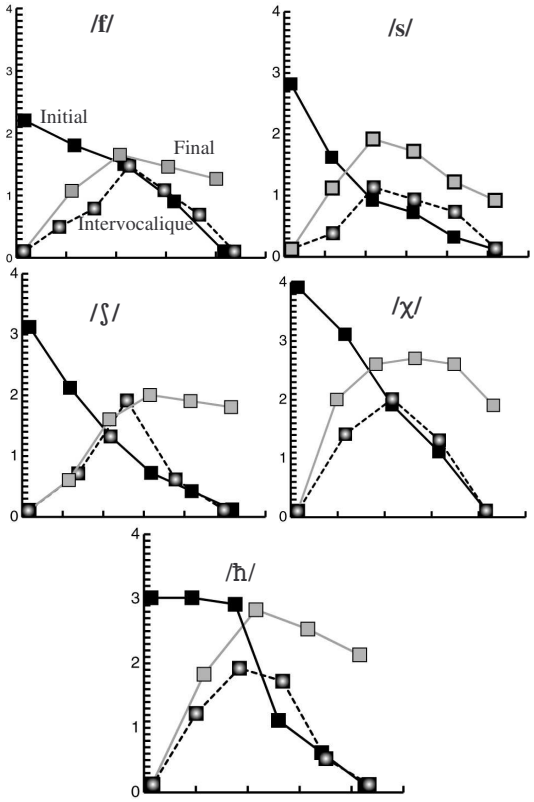


Figure 2. *Degré et durée d'ouverture glottale des fricatives sourdes dans les trois postions.*

3.2.1 Position initiale

La comparaison entre la configuration glottale des différents points d'articulation a révélé quelques différences importantes. Concernant le degré d'ouverture glottale, l'observation des différentes sessions des films fibroscopiques montre que l'ouverture de la glotte augmente à mesure que le lieu d'articulation recule dans la cavité buccale (ainsi f < s < ʃ < χ < ħ). En outre, l'uvulaire /χ/ est produite avec une vélocité de la fermeture glottale plus rapide. La posture du larynx pendant la réalisation de la pharyngale /ħ/ est, quant à elle, très marquée. La production de ce segment nécessite systématiquement un rapprochement important entre les sommets des aryténoïdes et le tubercule de l'épiglotte. Notons que pendant la tenue de ce segment, les aryténoïdes demeurent largement écartées.

3.2.2 Position intervocalique

Les mêmes observations concernant les fricatives initiales s'appliquent plus ou moins pour la position intervocalique. En effet, dans cette position aussi, les fricatives postérieures sont réalisées avec une ouverture glottale plus large comparées aux fricatives antérieures (même si cette différence n'est pas aussi importante et systématique qu'en position initiale). /ħ/ se réalise, comme en position initiale, avec un rapprochement très important entre les sommets des aryténoïdes et la base de l'épiglotte. Ce rapprochement est plus important en position intervocalique qu'en position initiale. En effet, à environ deux images avant l'onset de la consonne, le locuteur entame déjà un rapprochement entre les sommets des aryténoïdes et la base de l'épiglotte. Ce rapprochement se maintient aussi pendant la première moitié de la voyelle qui suit.

3.2.3 Position finale

Comme en position initiale, l'amplitude de l'ouverture glottale augmente à mesure que le lieu d'articulation recule dans la cavité buccale. Paradoxalement, l'état de la glotte après l'offset des fricatives, loin de s'ouvrir encore d'avantage, semble d'abord entamer un léger geste de fermeture et de rapprochement des aryténoïdes. Ce n'est qu'une ou deux images après que la glotte reprend la configuration qu'elle a pendant la phase respiratoire. Comme pour les positions initiale et intervocalique, l'uvulaire /χ/ se caractérise par une vélocité d'ouverture glottale plus rapide. Ainsi, et ce juste une image après la voyelle, la glotte atteint déjà un degré d'ouverture supérieur ou égal à l'ouverture maximale atteinte durant la tenue de /f/, /s/ ou /ʃ/. La pharyngale /ħ/ est produite, comme dans les autres positions, avec une compression antérieure-postérieure au niveau du sphincter aryépiglottique.

4. DISCUSSION

Des différences notables ont été observées selon le point d'articulation des obstruentes sourdes dans les trois positions. Concernant les occlusives sourdes, l'uvulaire présente la plus large amplitude glottale et la dentale emphatique, la plus faible. C'est principalement ce faible degré d'ouverture au moment du relâchement qui explique la durée plus courte du VOT pour les dentales emphatiques (voir [1]). La dentale et la vélaire sont produites avec une amplitude intermédiaire et présentent pratiquement la même ouverture maximale de la glotte, et la même amplitude au moment du relâchement oral. L'absence de différences entre /t/ et /k/ n'est pas propre aux segments berbères.

Löfqvist [2] et Zeroual [3] n'ont pas non plus relevé de différence d'amplitude glottale entre ces deux points d'articulation en suédois et en arabe marocain, respectivement. Selon Löfqvist, les différences relevées dans certaines langues, comme l'islandais et le danois (où les vélaires présentent une plus large amplitude), sont probablement dues à l'influence des mouvements verticaux du larynx. Les différences entre l'uvulaire d'un côté et les autres occlusives de l'autre, semblent liées aux différences manifestes dans la durée totale de l'ouverture glottale. Comme le montre la figure 1, /q/ est plus long que /t/, /k/ et /t/, de sorte que plus la durée de l'ouverture glottale est longue, plus l'amplitude de cette ouverture est large. L'uvulaire se comporte dans ce sens comme les occlusives géminées sourdes du berbère. Il a été démontré en effet dans un travail antérieur [1], que les occlusives géminées sourdes du berbère sont systématiquement produites avec une ouverture glottale maximale plus importante que leurs contreparties simples.

Parmi les fricatives, les consonnes /χ/ et /ħ/ sont produites avec des ajustements qui les distinguent clairement des autres segments. L'uvulaire est systématiquement produite avec un degré d'ouverture glottale plus important que les fricatives antérieures, et ce dans toutes les positions. Cela démontre, comme le confirment nos données, que le degré de l'ouverture glottale augmente à mesure que le lieu d'articulation recule dans la cavité buccale L'uvulaire entame aussi sa phase d'abduction plus rapidement. Ces deux aspects, observés aussi en arabe marocain (voir [3]), sont étroitement liés. La glotte, s'ouvrant plus rapidement, atteint tout naturellement une amplitude glottale plus importante plus rapidement. Ces ajustements glottaux sont probablement produits pour satisfaire des exigences d'ordre aérodynamique. Pendant la production d'une fricative dentale, un bruit de friction peut être produit sans exécuter une très large ouverture glottale. La raison en est que pendant la production

de ce segment, la constriction orale est assez étroite, un débit d'air faible suffit donc pour produire le bruit de friction nécessaire. Pendant la production de /χ/, la constriction supraglottique étant moins étroite, il faut un débit d'air plus important pour produire la turbulence nécessaire ([4], [5]).

La posture globale du larynx pendant la tenue de /ħ/ soulève la question de son lieu d'articulation : s'agit-il réellement d'une pharyngale, comme le laisse entendre les différentes descriptions du système consonantique berbère ou plutôt d'une aryépiglottale ? Ladefoged et Maddieson ([6] : 167) se posent la même question concernant deux autres langues supposées avoir ce type de consonnes (l'arabe et l'hébreu). Pour eux : « […] pharyngeal fricatives are not as common as might be supposed from the literature, as most of the sounds to which this label is attached (e.g. in Arabic and Hebrew) are actually what we would call epiglottal rather than pharyngeal in place.» Notons tout d'abord, que ce segment est un emprunt ancien à l'arabe. Dans cette langue aussi, et selon les observations effectuées par Zeroual [3], cette consonne présente les caractéristiques d'une aryépiglottale et non d'une pharyngale. Sur la base de nos observations fibroscopiques, et en comparaison avec les données de l'arabe marocain, langue avec laquelle le berbère est en contact étroit depuis plusieurs siècles, il semble plus approprié d'appeler cette consonne une aryépiglottale et non une pharyngale et de la transcrire, selon l'API, comme /ʜ/ au lieu de /ħ/.

Il a été largement observé que l'amplitude de l'ouverture glottale est plus importante pour les fricatives que pour les occlusives sourdes. C'est le cas par exemple en anglais, en islandais, en danois, en allemand, et en japonais (voir [7] pour une revue). A l'exception du cas de /q/ traité ci-dessus, la même caractéristique est observée en berbère comme le montre la mise en parallèle des figures 1 et 2. Une des raisons les plus avancées pour expliquer cette tendance, qui semble être universelle, est d'ordre aérodynamique. Pour Löfqvist & Yoshioka ([8] : 800) : "The difference in laryngeal movements between stops and fricatives [...] is most likely related to different aerodynamic requirements for stop and fricative production. A rapid increase in glottal area would allow for the high air flow necessary to generate the turbulent noise source during voiceless fricatives". Une autre différence entre les fricatives et les occlusives, attestée dans plusieurs langues, concerne la vélocité du geste d'ouverture glottale qui est plus importante pour les fricatives. Nos données confirment cet aspect aussi bien en position intervocalique, qu'en positions initiale et finale. On observe aussi que les cordes vocales continuent de vibrer avec une ouverture glottale plus importante pour /s/ que pour /t/ au moment de l'implosion. La rapidité de l'abduction des cordes vocales pour les fricatives peut expliquer cet aspect.

On peut en effet supposer que la glotte, s'ouvrant plus rapidement, atteint une amplitude assez importante avant que la différence de pression transglottique diminue à un niveau propice à la cessation des vibrations des cordes vocales. Cette vélocité du geste d'ouverture glottale peut aussi expliquer un aspect largement attesté : durant la tenue d'une séquence d'obstruantes sourdes, la glotte atteint généralement son niveau d'ouverture maximale pendant la tenue des fricatives (voir Ridouane et al. [9] et les références qu'ils citent).

RÉFÉRENCES

[1] Ridouane 2003. Suites de consonnes en berbère chleuh : phonétique et phonologie. Thèse de Doctorat Unifié, Université Paris 3.

[2] Löfqvist (1976). Closure duration and aspiration for Swedish stops. *Phonetics Laboratory Working Papers* 13, 1-39. Lund University.

[3] Zeroual, C. (2000). Propos controversés sur la phonétique et la phonologie de l'arabe marocain. Thèse de Doctorat Unifié, Université Paris 8.

[4] Yeou, M. & Maeda, S. (1995). Pharyngeal and uvular consonants are approximants: An acoustic modeling study. Proceedings of the 13th International Congress of Phonetic Sciences, 586-589.

[5] Stevens, K.N. (1998). *Acoustic phonetics*. Cambridge MA, London.

[6] Ladefoged, P. & Maddieson, I. (1996). The sounds of the world's languages. Blackwell Publishers: Oxford.

[7] Yoshioka, H., Löfqvist, A. & Hirose H. (1980). Laryngeal adjustments in Japanese voiceless sound production. Haskins Laboratories: Status Report on Speech Research SR-63/64, 293-308.

[8] Löfqvist, A. & Yoshioka, H. (1980). Laryngeal activity in Swedish obstruent clusters. Journal of the Acoustical Society of America 68(3), 792-801.

[9] Ridouane, R., Fuchs, S., & Hoole, P. (2006). Laryngeal adjustments in the production of voiceless obstruent clusters in Berber. In Harrington & Tabain (eds.). Speech Production: Models, Phonetic Processes, and Techniques, 249-267. Macquarie University: Sydney, Australia.

Extraction semi-automatique des mouvements du conduit vocal à partir de données cinéradiographiques

Julie Fontecave et Frédéric Berthommier

ICP – Institut de la Communication Parlée
INPG, 46 avenue Félix Viallet, 38000 Grenoble, France
fonte,bertho@icp.inpg.fr

ABSTRACT

Since high speed X-ray films still provide the best dynamic view of the entire vocal tract, large existing databases have been preserved and are available for the speech research community. We propose a new technique for automatic extraction of vocal tract movements from these data. At first, the method was developed for the extraction of tongue movements in Wioland recorded in Strasbourg in 1977. Then, the same technique was adapted to other articulators and other X-rays films, taking into account their specificities. Finally, a quantitative evaluation of the estimate error and a comparison with Thimm and Luettin (1999) are achieved.

1. INTRODUCTION

La radiographie a été pendant longtemps l'une des principales techniques d'acquisition de données articulatoires en offrant la possibilité d'obtenir une vue sagittale complète des articulateurs du conduit vocal, de la glotte jusqu'aux lèvres. Devenue dynamique à la fin des années 1950, sous le terme de cinéradiographie, elle permet l'observation des mouvements des articulateurs de la parole avec une résolution temporelle importante, de l'ordre de 60 ips (images par seconde). Depuis quelques années, pour des questions de déontologie, on n'enregistre plus de nouveaux films radiologiques du conduit vocal. La cinéradiographie ayant fait la preuve de son utilité pour la recherche scientifique, il est nécessaire de pouvoir continuer à utiliser les données existantes en préservant les films. C'est dans ce contexte que Munhall et coll. [1] ont réalisé la base ATR « X-ray film database for Speech Research », à partir de films enregistrés par Rochette (Université Laval), et Stevens et Perkell (M.I.T.). Soutenu par le programme « Ingénierie des Langues » du CNRS, l'Institut de Phonétique de Strasbourg et l'Institut de la Communication Parlée de Grenoble ont aussi élaboré une base de données cinéradiographiques du français incluant les séquences Wioland et Flament [2].

L'extraction de données géométriques à partir de films radiologiques est généralement réalisée manuellement, mais on doit faire face à de grandes quantités de données pour traiter la moindre séquence. L'extraction automatique des contours de la langue fût envisagée par Laprie et Berger [3] pour exploiter au mieux ces grandes bases. Mais jusqu'à présent, seuls les travaux de Thimm et Luettin [4] ont aboutis au traitement complet d'un film issu de la base ATR (Laval43).

En vue d'améliorer cette situation, nous avons mis en place une méthode semi-automatique applicable film par film et qui combine le marquage manuel et la reconstruction automatique du mouvement. Cette technique [5] est basée sur une adaptation de l'algorithme de rétro-marquage [6], dont le principe est d'associer des paramètres implicites et extraits du signal vidéo à des paramètres géométriques contrôlés et définis a posteriori, plutôt que d'extraire directement des données géométriques. Pour estimer les mouvements de langue, la méthode se décompose en 3 étapes : (1) le traitement manuel d'un nombre restreint d'images clefs qui permet de définir des paramètres géométriques (ici le contour de la langue), (2) une étape automatique d'indexation de la base à partir de ces mêmes images clefs réduites et cadrées, qui a pour but d'associer à chacune des images de la base le marquage géométrique et (3) des traitements postérieurs de régularisation. A noter que le rétro-marquage peut être rendu entièrement automatique lorsque les informations géométriques sont extractibles dans les images clefs (voir un exemple, dans http://www.icp.inpg.fr/~bertho/m2p/jep06/main.wmv, sur les mouvements de la main). Mais dans le cas de la langue, cette tâche très difficile même pour l'expert humain est dévolue au marquage manuel dans des conditions de facilitation que nous décrirons par la suite.

A l'heure actuelle, cette méthode a aisément été appliquée avec succès sur plusieurs films radiographiques et adaptée pour tirer profit des particularités de ces bases. D'abord sur Wioland [7] pour la mise au point, puis sur le film Flament [8], composé de près de 5000 images, avec lequel nous nous intéressons plus particulièrement à la pointe de la langue et au voile du palais. Enfin, l'application de la méthode sur l'une des séquences de la base de données d'ATR, Laval43, a permis la comparaison directe de nos résultats avec ceux de Thimm et Luettin [4] cités précédemment.

2. MÉTHODE QUASI-AUTOMATIQUE D'EXTRACTION DE MOUVEMENTS

2.1. Extraction des mouvements de la langue à partir de la base Wioland

Cette séquence enregistrée en 1977 [7] et numérisée récemment comprend 5673 images du conduit vocal

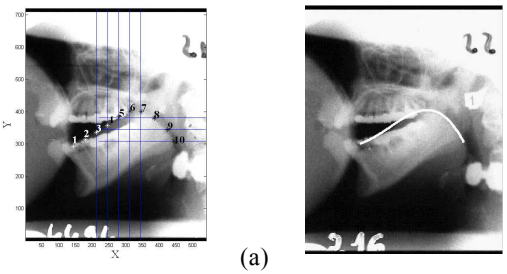
provenant de 64 séquences vidéos (64 phrases prononcées par une locutrice française), enregistrées à 66 ips. Durant la phase de mise au point, notre méthode a d'abord permis de récupérer les mouvements de la langue, puis elle a aisément pu être étendue à d'autres parties du conduit vocal (lèvres, vélum…).

L'étape manuelle consiste à décrire, pour 100 images clefs choisies aléatoirement, la position du contour de la langue avec 10 points (Fig. 1a), dont 8 par intersection avec des lignes verticales et horizontales (base et dos) et 2 points libres pour la pointe, soit 12 degrés de liberté (ddl). Cette étape est réalisée avec une interface qui permet, grâce à un curseur actionné manuellement, de voir la langue en mouvement, et dans de nombreux cas, d'associer un contour quasiment indiscernable sur l'image clef statique. Le choix des lignes de marquage et des points libres est fait de telle sorte qu'il n'y ait pas de données manquantes. A ce stade, pour chaque image clef, on dispose d'une configuration géométrique brute pour la langue en reliant les 10 points.

Puis, pour chaque image de la séquence, l'index de l'image clef la plus proche est assigné. La mesure de similarité est la distance Euclidienne entre les coefficients DCT (Discrete Cosinus Transform) basses fréquences des deux images. Au préalable, les images sont restreintes à un cadre minimal d'observation de l'articulateur cible pour tout le film, ceci de façon à minimiser l'interférence avec d'autres articulateurs ou des parasites (e.g. les inscriptions manuelles visibles figure 1). Après indexation, on aboutit ainsi à un premier marquage de la base entière en associant l'information géométrique disponible uniquement pour les images clefs. Des traitements postérieurs, filtrage temporel et moyennage de configurations voisines obtenues par multi-indexation permettent d'améliorer sensiblement cette première estimation entachée d'une erreur de quantification et des erreurs dues à l'indexation. Un lissage spline est également appliqué sur les points estimés. Nous aboutissons ainsi à une reconstruction complète du mouvement observable par superposition sur le film d'origine (Fig. 1b).

En terme de temps de traitement, le marquage manuel de 10 points sur 100 images clefs est estimé à 2 heures minimum. Ensuite, le temps d'exécution de la méthode pour la base complète (5673 images) dure quelques minutes.

Une évaluation objective a été mise en place à l'aide d'un deuxième jeu de 100 images tests marquées. L'erreur RMS par ddl est calculée entre les marques manuelles et les marques estimées par la méthode quasi-automatique. L'erreur $Etot_1$ considérée au final est la moyenne de cette erreur sur les 12 degrés de liberté. Elle est égale à 11 pixels (à comparer avec 350 pixels de longueur totale) équivalents à 3 mm après une calibration approximative car cette information n'est pas disponible directement sur le film.



(a) (b)

Figure 1 : (a) Excepté pour la pointe, les points sont marqués à l'intersection entre le contour de la langue et les lignes verticales ou horizontales.
(b) Le résultat est observé sur une vidéo (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/langue-wioland.wmv) par superposition des configurations géométriques estimées.

2.2. Estimation des mouvements du conduit vocal

Nous avons appliqué la même méthode aux autres parties du conduit vocal, essentiellement lèvres et vélum. Pour chaque articulateur, les images d'origine sont découpées de façon à inclure l'élément à marquer pour tout le film, et à exclure les interférences. Les paramètres de la méthode (nombre d'images clefs, degrés de liberté, nombre de coefficients DCT nécessaires pour l'indexation) sont définis de façon indépendante pour chaque articulateur. Les parties fixes du conduit vocal (palais, pharynx) sont également marquées de façon à reconstruire les mouvements du conduit vocal complet (figure 2).

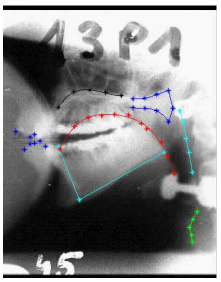


Figure 2 : Marquage complet du conduit vocal dans Wioland (http://www.icp.inpg.fr/~bertho/m2p/jep06/conduit-wioland.wmv)

Les mouvements du conduit vocal pourront être associés avec l'audio afin d'étudier les aspects dynamiques de la relation entre configuration géométrique du conduit vocal et acoustique.

3. BASE CINÉRADIOGRAPHIQUE FLAMENT : POINTE ET VOILE DU PALAIS

Pour traiter la base cinéradiographique Flament, enregistrée dans des conditions proches de Wioland, 13 ddl ont été définis pour marquer le contour de la langue : 9 points à 1 ddl pour la base et le dos et 2 points à 2 ddl pour la pointe. La pointe de la langue est nettement plus

visible dans ce film et une adaptation a été réalisée afin de mieux capturer ses mouvements rapides et parfois relativement indépendants. Elle consiste en un double marquage associant une estimation globale des 13 ddl comme précédemment, et une seconde spécifique de la pointe incluant 5 ddl seulement. Cette dernière est calculée à partir d'un cadre focalisé sur la pointe (Fig. 3). Le nombre d'images clefs a aussi du être étendu à 200. La fusion de ces deux estimations est réalisée par substitution des 5 ddl de la pointe dans l'estimation globale (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/langue-flament.wmv). L'erreur de reconstruction $Etot_1$ est estimée à 10 pixels pour une longueur moyenne de langue de 375 pixels.

Le suivi des mouvements de la pointe permet en particulier de détecter les instants de contact de la langue, que l'on peut rapprocher de l'audio et des événements consonantiques.

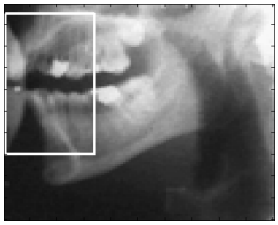


Figure 3 : La position de la pointe est estimée localement à partir d'un cadre spécifique.

D'autre part, le corpus est dédié à la question des nasales du Français [8] et le vélum est bien visible sur le film. Un traitement spécifique à cet articulateur a aussi été réalisé avec succès (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/velum-flament.wmv).

4. Base cinéradiographique d'ATR : un comparatif sur la langue

La base de données d'ATR est la plus grande disponible pour les recherches en parole, avec 25 films différents totalisant une durée de 55 minutes et près de 100000 images. Nous avons extrait ces images à partir du DVD mais il n'est pas pour l'instant possible d'en réaliser le traitement complet à cause de l'étape de marquage manuel propre à chaque film. Notre étude concerne la séquence Laval43 dans un but comparatif.

4.1. Méthode d'extraction développée à l'IDIAP

Le film Laval43 (environ 4000 images) a été marqué en totalité par Thimm et Luettin [4] à l'IDIAP. En résumé, les résultats, disponibles en détail sur le site http://www.idiap.ch/machine_learning.php?project=64 ont été obtenus à partir d'une technique de normalisation d'histogrammes et d'une méthode d'extraction de contours. La méthode, notée TL, fait aussi appel à des images clefs choisies aléatoirement, sur lesquelles l'application d'un détecteur de Canny permet de récupérer le contour de la langue. La procédure de suivi de contours utilise l'appariement avec ces images clefs et l'information temporelle. Les résultats concernent plusieurs articulateurs du conduit vocal, mais ne permettent pas de reconstruire sa forme complète, en particulier car la pointe de la langue est souvent manquante. Nous nous intéresserons ici aux résultats concernant la langue, dans le but de comparer directement cette méthode à la nôtre, notée FB. Ces résultats ont été récupérés et conditionnés, de même que le film Laval43, pour permettre une comparaison objective. Nous disposons ainsi d'un jeu de splines définissant le contour de la langue pour chaque image. Nous le noterons S_{TLi} dans la suite de l'article.

4.2. Comparaison de méthodes

La méthode de rétro-marquage a été appliquée avec 200 images clefs, 9 points à 1 ddl et 2 points à 2 ddl pour la pointe, soit 13 ddl. Nous disposons ainsi d'un second jeu de splines, noté S_{FBi}.

Pour comparer les 2 estimations à partir de ces 2 jeux de splines, 2 types de mesures sont considérés :

- une mesure relative D, qui calcule la distance entre les 2 splines, proposée par Thimm [9]. Il s'agit de l'aire comprise entre les 2 courbes splines, normalisée par la somme de leurs longueurs.

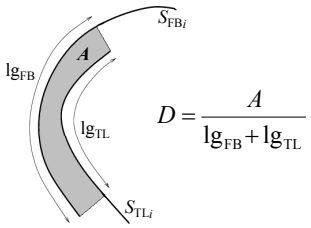


Figure 4 : Mesure de distance entre 2 splines

- une mesure de $Etot_1$ basée sur des images tests mesurant l'écart entre le marquage manuel et les 2 estimations (Fig. 5). Pour cette mesure, compte-tenu des données manquantes pour la pointe (dues à la difficulté d'estimation par une approche contour), nous n'avons pris en compte que les 8 points définissant le contour du dos et de la base (Fig. 5).

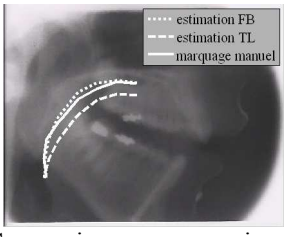


Figure 5 : Comparaison sur une image test d'un marquage manuel de la langue avec 2 marquages estimés (la pointe est exclue de cette comparaison)

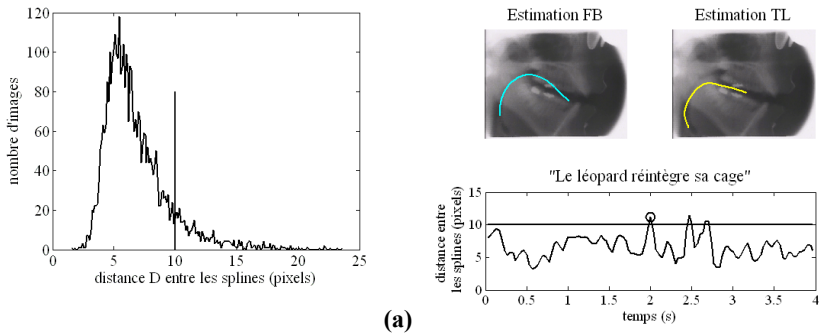


Figure 6 : (a) Répartition de la différence *D* entre les 2 estimations, et définition d'un seuil de décrochage *D*>10
(b) Décrochage observé au milieu de la séquence considérée, avec les deux contours estimés à cet instant

La différence moyenne *D* entre les 2 estimations est de 6.8 pixels. La distribution de *D* (Fig. 6a) montre que pour 10% de la base, il existe un décrochage entre les 2 méthodes que nous caractérisons par un seuil $D > 10$ pixels. Au dessus du seuil, l'écart moyen est de 12.7 pixels. Ce décrochage est observable visuellement sur la figure 6b. A noter que le contour associé par rétro-marquage est correct dans ce cas, et qu'il inclut la pointe.

Avec la méthode de rétro-marquage, l'erreur $Etot_1$ calculée sur 8 ddl varie en fonction du nombre d'images clefs et des traitements postérieurs (Fig. 7). Nous obtenons ainsi une erreur inférieure à 8 pixels (condition 175 clefs, 4v af) et autour de 20 pixels pour Thimm et Luettin (1999), sachant que cette section de la langue a une longueur estimée de 250 pixels.

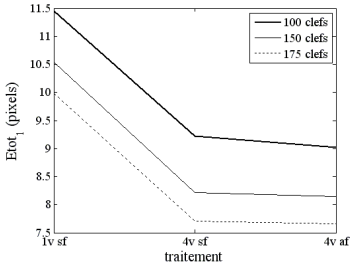


Figure 7 : Evolution de l'erreur $Etot_1$ pour notre méthode de marquage avec différents traitements postérieurs (indexation simple ou multiple, sans ou avec filtrage)

5. CONCLUSION

Nous montrons que le rétro-marquage basé sur les paramètres vidéo DCT basses fréquences offre la possibilité de suivre le mouvement du tractus vocal, même lorsque les contours ne sont pas entièrement visibles, notamment pour la pointe. C'est un progrès, mais la question reste ouverte de savoir si nos mesures sont suffisamment précises pour être mises en correspondance avec les caractéristiques temporelles et spectrales de la parole afin d'en obtenir de nouvelles informations. Par contre, en ce qui concerne le mouvement du velum, la cinéradiographie apportera sans doute des données précieuses qui ne sont pas accessibles autrement.

REMERCIEMENTS

Nous remercions Pascal Perrier pour les films cinéradiographiques Wioland et Flament, numérisés dans le cadre du programme « Ingénierie des Langues » du CNRS. Nous remercions Kevin Munhall pour nous avoir adressé une version de la base ATR sur DVD.

BIBLIOGRAPHIE

[1] K.G. Munhall, E. Vatikiotis-Bateson & Y. Tohkura. X-ray Film database for speech research. *Journal of the Acoustical Society of America,* 98 : 1222-1224, 1995.

[2] A. Arnal, P. Badin, G. Brock, P.-Y. Connan, E. Florig, N. Perez, P. Perrier, P. Simon, R. Sock, L. Varin, B. Vaxelaire & J.-P. Zerling. Une base de données cinéradiographiques du français. *XXIIIèmes Journées d'Etude sur la Parole*, pages 425-428, 2000.

[3] Y. Laprie & M.-O. Berger. Extraction of Tongue Contours in X-Ray Images with Minimal User Interaction. In *Proc. Int. Conf. on Spoken Language Processing,* volume 1, pages 268-271 , 1996.

[4] G. Thimm & J. Luettin. Extraction of articulators in X-ray image sequences. In *Proc. Eur. Conf. on Speech Communication and Technology*, pages 157-160, 1999.

[5] J. Fontecave & F. Berthommier. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database. In *Proc. Eur. Conf. on Speech Communication and Technology*, pages 1081-1084, 2005.

[6] F. Berthommier. Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 3, pages 789-792, 2004.

[7] F. Wioland. Faits de jointure en français. Implications aux niveaux articulatoire et acoustique. Incidences sur le plan des fonctions linguistiques. *Doctorat d'Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg , 1985.

[8] B. Flament. Recherche sur la mise en relief en français. Approche théorique et essai de caractérisation phonétique à partir de données de la mingographie et de la radiocinématographie. *Doctorat d'Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg, 1984.

[9] G. Thimm. Segmentation of X-ray image sequences showing the vocal tract. *IDIAP Research Report*, IDIAP, Suisse, 1999.

Session XVII

Analyse, codage et synthèse

Jeudi 15 juin 2006 - 14h30 16h30

Bases théoriques et expérimentales pour une nouvelle méthode de séparation des composantes pseudo-harmoniques et bruitées de la parole

Laurent GIRIN

Institut de la Communication Parlée, INPG/Univ. Stendhal/CNRS UMR 5009
46 av. Félix Viallet, 38040 Grenoble, France
Tél: +33 476 57 47 15, fax: +33 476 57 47 10, email : girin@icp.inpg.fr
Web : http://www.icp.inpg.fr/~girin/

ABSTRACT

In this paper, the problem of separating the harmonic and noise components of speech signals is addressed. A new method is proposed, based on two specific processes dedicated to better take into account the non-stationarity of speech signals: first, a period-scaled synchronous analysis of spectral parameters (amplitudes and phases) is done, referring to the Fourier series expansion of the signal, as opposed here to the typically used Short-Term Fourier Transform (STFT). Second, the separation itself is based on a low-pass filtering of the parameters trajectory. Preliminary experiments on synthetic speech show that the proposed method has the potential to significantly outperform a reference method based on STFT: Signal-to-error ratio gains of 5 dB are typically obtained. Conditions to go beyond the theoretical framework towards more practical applications on real speech signals are discussed.

1. INTRODUCTION

Les composantes du signal de parole peuvent être grossièrement classifiées en deux catégories, selon la nature de la source vocale : d'un côté les composantes *harmoniques* (H) sont générées par les vibrations des cordes vocales, et d'un autre côté, les composantes *bruitées* (B) sont générées par une source de bruit fricatif, plosif ou d'aspiration [1]. Comme les sources H/B peuvent être simultanées, ces composantes sont souvent mélangées au niveau de la réalisation acoustique du signal. Pour un son donné, la contribution respective de ces composantes peut être quantifiée par l'estimation d'un rapport (de puissance) harmonique à bruit (RHB) (voir une revue dans [2]). Plus difficile, la *séparation* complète des composantes harmoniques et bruitées du signal mixte est un challenge important dans un certain nombre d'applications de traitement de la parole [3]–[5] (et aussi de la musique [6]). L'objectif est d'obtenir deux signaux à partir du signal original : un signal estimé complètement voisé et un autre complètement non-voisé, tels que la somme des deux soit égale au signal original. Ces deux signaux peuvent alors être séparément analysés, modélisés, et modifiés, en particulier pour la synthèse [7], le codage [8], et l'étude fondamentale de la production de parole.

Plusieurs méthodes ont été proposées dans la littérature pour l'estimation du RHB [2], et pour la séparation H/B [3]–[6]. Les méthodes en fréquence sont quasiment toutes basées sur la transformée de Fourier à court-terme (TFCT) pour l'analyse et la synthèse : grosso modo, les pics dominants du spectre sont supposés correspondre aux harmoniques, et les régions du spectre « irrégulières » ou entre les pics sont supposés correspondre aux composantes de bruit. Une telle approche est limitée par un facteur crucial : la parole est un signal localement quasi-stationnaire, et non strictement stationnaire. Ceci signifie que les composantes harmoniques et bruitées évoluent continûment dans le temps, plus ou moins lentement, et c'est une difficulté majeure pour une méthode précise de séparation H/B de ne pas considérer l'évolution des harmoniques d'une période à la suivante comme une partie des composantes bruitées [2]. Pourtant, dans la littérature, les fenêtres d'analyse-synthèse comprennent plusieurs périodes de signal, et la TFCT est intrinsèquement un processus moyenneur qui ne capture pas les différences précises entre ces périodes mais qui extrait plutôt leur caractéristiques communes pour les identifier à des composantes harmoniques constantes sur la fenêtre.

Dans cette étude, nous posons les bases d'une nouvelle méthode de séparation H/B travaillant à l'échelle de la période de signal, de façon à suivre précisément l'évolution des paramètres du signal d'une période à l'autre. C'est une méthode travaillant à la fois en temps et en fréquence, car elle se fonde sur la décomposition en séries de Fourier de chaque période de signal (au lieu de l'habituelle TFCT). La séparation résulte d'un filtrage des trajectoires des paramètres analysés. C'est pourquoi la méthode s'appelle Séparation H/B par filtrage des trajectoires de paramètres spectraux période-synchrones (FTPS2).

Cet article est organisé de la façon suivante. La méthode de séparation H/B est présentée en Section 2. La méthodologie de test est donnée à la Section 3. Résultats et perspectives sont données aux Sections 4 et 5.

2. LA METHODE FTPS2

2.1. Principe général

Soit un signal mixte voisé / non voisé, comprenant un grand nombre K de (pseudo-)périodes $s_k(n)$. Chacune de ces K périodes est décomposée au sens des séries de Fourier réelles, comme une somme de cosinus harmoniques :

$$s_k(n) = \sum_{i=1}^{I} A_i^k \cos(i\omega_0^k n + \theta_i^k) \quad k = 1 \text{ à } K \qquad (1)$$

Le signal complet est donc représenté par I jeux de K amplitudes A_i^k et phases relatives θ_i^k ($i = 1$ à I, $k = 1$ à K), plus un jeu de la fréquence fondamentale ω_0^k, $k = 1$ à K. Pour un signal pseudo-périodique, l'évolution des amplitudes et des phases d'une période à la suivante doit être plutôt lente ou « lisse », à cause de la nature

déterministe du signal. Au contraire, les composantes apériodiques/bruitées ont une nature aléatoire, et les paramètres spectraux associés (en particulier les phases) doivent varier de façon beaucoup plus importante [8]. Puisque les paramètres sont extraits sur un signal mixte H/B, leur évolution prend typiquement la forme d'une trajectoire de fond lente/lisse, supposée due aux composantes pseudo-harmoniques, à laquelle se superpose un bruit de type additif, supposé dû aux composantes bruitées. Par conséquent, l'extraction du signal harmonique à partir du signal mixte se fait en extrayant la trajectoire de fond lisse des paramètres par un filtrage passe-bas, et en l'identifiant à celle des paramètres des composantes harmoniques. Le signal harmonique estimé est alors généré en appliquant (1) avec les paramètres filtrés à la place des paramètres initiaux. Finalement, le signal de bruit est estimé en soustrayant le signal harmonique estimé au signal mixte. Il est important de noter que cette technique est en fait équivalente à un moyennage *glissant* qui respecte la dynamique des paramètres *à l'échelle de la période*. Ceci s'oppose au moyennage brut de la TFCT déjà mentionné. On cherche dans cette étude à retrouver la « vraie » trajectoire des paramètres harmoniques à partir de mesures perturbées par les composantes bruitées, et à l'inverse de la resynthèse par TFCT inverse, la méthode proposée garantit de reconstruire un signal harmonique estimé qui évolue d'une période à l'autre.

2.2. Détails techniques

Analyse des paramètres : la méthode proposée suppose que le signal à traiter est d'abord segmenté en périodes successives. Dans cette étude, les tests portent sur des signaux synthétiques (voir Section 3.1). Les frontières de périodes (*pitch-marks*) sont donc exactement contrôlées et utilisées dans le processus d'analyse. Dans l'optique d'une extension future aux signaux réels, différentes méthodes peuvent être utilisées pour estimer automatiquement les *pitch-marks*. Clairement, la précision de la méthode de séparation H/B dépend fortement de celle de cette estimation. Nous n'insistons pas sur ce point dans cet article car nous focalisons sur les bases théoriques de la méthode et sur de premières confirmations expérimentales du bien-fondé de ces bases. Cependant, des solutions pour dépasser cette difficulté sont proposées à la Section 5. Ainsi, dans cette étude, ω_0^k est donnée directement par l'inverse de la longueur de la période k. Puis, étant donnée ω_0^k, les amplitudes A_i^k et les phases θ_i^k sont estimées en utilisant la procédure de George et Smith de [9]. Cette procédure est une minimisation itérative au sens des moindres carrés de l'erreur entre le modèle harmonique de (1) et le signal. Elle permet d'obtenir une estimation précise des paramètres avec un très faible coût de calcul.

Régularisation de la phase: les mesures de phase sont obtenues modulo 2π. Il faut donc d'abord s'assurer qu'aucun saut de 2π artificiel ne vient perturber leur trajectoire « naturelle ». Pour cela, une procédure de régularisation de ces mesures le long de l'axe temporel est appliquée. Elle consiste à ajouter ou retrancher itérativement 2π à chaque mesure de phase si cela permet de diminuer la variance du vecteur compilant les mesures. Comme, la trajectoire de fond évolue au cours du temps, la variance est calculée avec une fenêtre glissante (typiquement, quatre périodes peuvent être utilisées) et plusieurs passes peuvent être effectuée. A la fin de cette procédure, les mesures sont parfaitement homogènes, bien que toujours bruitées.

Filtrage des paramètres : comme expliqué en Section 2.1, l'étape suivante composant le cœur de la méthode est le filtrage passe-bas des trajectoires des paramètres spectraux (amplitudes et phases). Des tests pilotes ont montrés qu'une large gamme de filtres très simples (*i.e.*, FIR avec peu de coefficients) fournissait des résultats assez proches. Dans les expériences de la Section 4, on utilise un filtre FIR à 10 coefficients avec une fréquence de coupure de 0.1 obtenu par la méthode de fenêtrage avec une fenêtre rectangulaire. Ce filtre est appliqué en mode *forward-backward* (filtrage à phase nulle), de façon à ce que les paramètres filtrés et non filtrés soient synchrones, et qu'il en soit de même pour les signaux H/B séparés et le signal mixte original.

Ré-estimation de l'amplitude : En pratique, on observe que les trajectoires des paramètres d'amplitude sont généralement plus bruitées que celles des paramètres de phase. Par conséquent, la méthode a été raffinée avec une ré-estimation des amplitudes après le filtrage des paramètres de phase. Ceci est fait avec une version simplifiée de la procédure d'analyse des moindres carrés itératifs utilisée précédemment, avec la phase maintenant fixée à la valeur obtenue après filtrage. Les amplitudes ainsi ré-estimées sont ensuite filtrées par le filtre passe-bas.

3. METHODOLOGIE DE TEST

3.1. Génération de signaux synthétiques

Des signaux synthétique mixtes voisés / non voisés sont générés de façon à ce que les « vraies » parties harmoniques et bruitées soient séparément disponibles. Ceci permet de calculer des mesures objectives de séparation, telles que le rapport signal à erreur (RSE) [2][4] que nous utilisons par la suite (voir la sous-section 3.3). Les signaux synthétiques consistent en différentes versions des voyelles /a/ d'une voix masculine et /i/ d'une voix féminine, prononcées de manière soutenue ($K = 300$) et échantillonnées à 48kHz (avec une bande passante limitée à 8 kHz). Leur génération suit la méthodologie utilisée dans [2][4]. Un train d'ondes glottales suivant le modèle de Rosenberg [10] est utilisé comme source harmonique. Un bruit blanc gaussien est utilisé comme source de bruit. Il est éventuellement modulé en amplitude par le train d'ondes glottales pour plus de naturel [7]. Les deux sources alimentent un filtre numérique tout-pôles modélisant le conduit vocal. Ce filtre résulte de l'analyse LPC à l'ordre 50 d'un signal réel produit par un locuteur masculin pour le /a/ et par une locutrice pour le /i/. Des filtres du premier ordre pour la pré-emphase et la simulation de radiation labiale sont aussi utilisés pour mieux caler le spectre du signal synthétique sur celui des signaux réels et permettre un son plus naturel. Le signal mixte est obtenu en sommant les deux signaux filtrés et centrés avec différents RHB dans la gamme –10 dB à 30 dB. Enfin, pour évaluer la robustesse de la méthode proposée sur des signaux non-stationnaires et plus proches du naturel, de la prosodie est générée par modulation de la fréquence fondamentale de la source glottale selon la formule :

$$\omega_0^k = \alpha + \beta \cos\left(2\pi \frac{3k}{K}\right) + \gamma \frac{k^2}{K^2} \qquad (2)$$

Le terme en cosinus assure trois cycles mélodiques et le terme quadratique assure une montée rapide à la fin de la voyelle. A la Section 4, on fournit des résultats pour une fréquence fondamentale fixe (*i.e.*, pour /a/, $\alpha = 130$, $\beta = \gamma = 0$; pour /i/, $\alpha = 280$, $\beta = \gamma = 0$), pour une intonation « normale » (*i.e.*, pour /a/, $\alpha = 130$, $\beta = 10$, $\gamma = 20$; pour /i/, $\alpha = 250$, $\beta = 10$, $\gamma = 20$), et pour une intonation « exagérée » (*i.e.*, pour /a/, $\alpha = 110$, $\beta = 30$, $\gamma = 100$; pour /i/, $\alpha = 200$, $\beta = 30$, $\gamma = 200$) (toutes les valeurs sont en Hz).

3.2. Une méthode de référence basée sur la TFCT

Pour évaluer comparativement notre méthode, nous avons implanté la méthode *Pitch-Scaled Harmonic Filter* (PSHF) de [4]. Cette méthode a été choisie car 1) elle est bien représentative des méthodes basées sur la TFCT 2) elle est relativement simple à implanter par rapport à d'autres méthodes [5] 3) son évaluation sur des signaux synthétiques a fourni des scores de RSE de référence. Son principe est de calculer des spectres successifs par TFCT sur exactement quatre périodes du signal mixte, de façon à ce que les pics des harmoniques soient supposés être localisés tous les quatre canaux spectraux de la TFCT et donc facilement isolés par un filtre peigne. Quatre périodes du signal harmonique estimé sont alors générées par TFCT inverse du spectre filtré par le peigne. Le signal harmonique complet est reconstruit par sommation pondérée des estimations successives. En le soustrayant au signal mixte, on obtient le signal de bruit estimé.

3.3. Mesures de Rapport Signal à Erreur (RSE)

L'évaluation objective de la séparation H/B est faite par calcul de rapport signal à erreur (RSE) qui peut être fait aussi bien pour le signal harmonique estimé que pour le signal de bruit estimé. Notons RSE_H le rapport de puissance entre le signal harmonique « vrai » et sa différence d'avec le signal harmonique estimé. De même, notons RSE_B le même rapport pour le signal de bruit. Comme le signal de bruit estimé est obtenu en soustrayant le signal harmonique estimé au signal mixte, on a : $RSE_H = RSE_B + RHB$. Par conséquent, par la suite nous ne considérons que RSE_B (noté simplement RSE), du fait qu'il s'est révélé quasiment constant en fonction du RHB dans [4].

4. RESULTATS

4.1. Rapports Signal à Erreur

La Figure 1 montre les RSE obtenus sur les voyelles de test, avec les deux méthodes, FTPS2 et PSHF, et pour les trois contours de ω_0. Les principaux résultats sont les suivants :

- Les deux méthodes fournissent des résultats remarquablement stables sur une large gamme de RHB : les RSE sont quasi-constants de –10 à environ 15 dB de RHB dans presque tous les cas. Pour la méthode PSHF, le RSE est d'environ 5 dB (de 5 à 5.4 dB dans la gamme de –10 à 15 dB de RHB, selon les conditions) et ce résultat est très cohérent avec ceux de [4], qui donnent une valeur stable typique de 5 dB.
- Les performances obtenues avec la méthode FTPS2 dépassent largement cette référence de 5 dB. Pour –10 à 15 dB de RHB, les valeurs de RSE sont toutes autour de 9.5 dB pour /a/ (sauf pour l'intonation exagérée) et autour de 10.5 dB pour /i/ (du moins pour l'intonation normale et exagérée; de façon surprenante, une valeur légèrement inférieure de 10 dB est obtenue pour ω_0 constant). Ainsi, la méthode FTPS2 à un gain de l'ordre de 4 à 5.5 dB par rapport à la méthode PSHF, selon les conditions. Un exemple typique de résultat est donné sur la Figure 2.
- Les performances des deux méthodes chutent pour un RHB supérieur à 15 dB, et plus la variation d'intonation est forte, plus la dégradation est forte. Ce résultat n'est pas surprenant : plus la partie bruitée du signal est faible, plus elle est difficile à séparer de la partie harmonique, et l'augmentation de la non-stationnarité du signal rend la tâche encore plus difficile. On peut aussi remarquer que les deux voyelles offrent une robustesse différente à ces dégradations, mais une discussion sur l'influence de facteurs phonétiques va au-delà du cadre de cet article. Notons que même dans des conditions difficiles, l'avantage de la méthode FTPS2 sur la méthode PSHF reste toujours supérieur à 4 dB, sauf pour /i/ avec intonation exagérée à 25–30 dB, où « seulement » 3.7 et 3.1 dB de gain sont obtenus. Pour les autres conditions, ce gain est typiquement de 5 dB, et il peut même aller au-delà, par exemple 8 dB pour /a/ avec ω_0 fixe à 30 dB.
- Finalement, on peut noter que les résultats obtenus avec ou sans la modulation de la source de bruit par la source glottale sont toujours très proches. C'est pourquoi on présente sur la Figure 1 seulement les résultats obtenus sans modulation. De façon complémentaire, sur l'exemple de la Figure 2, le signal de bruit est modulé. Ces résultats semblent indiquer que les deux méthodes sont assez robustes par rapport aux non-stationarités de la source de bruit. Ce point est important en regard de l'application de la séparation H/B sur des signaux réels et devra être étudié plus en détails dans le futur.

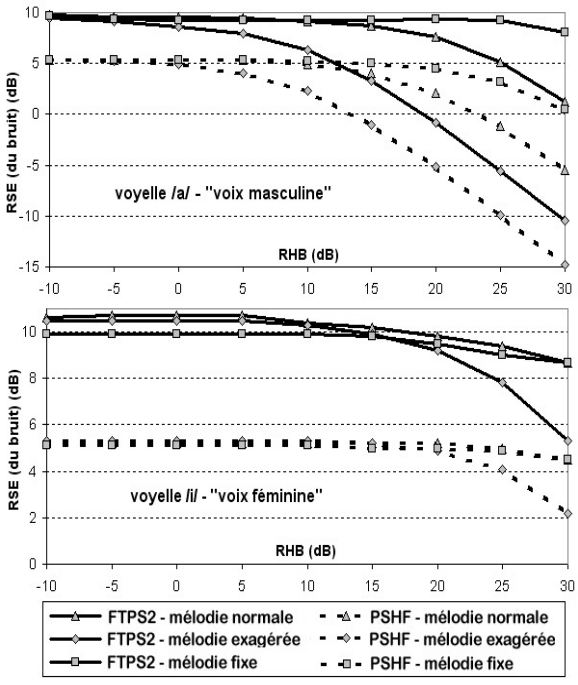


Figure 1 : RSE (du signal de bruit) en fonction du RHB.

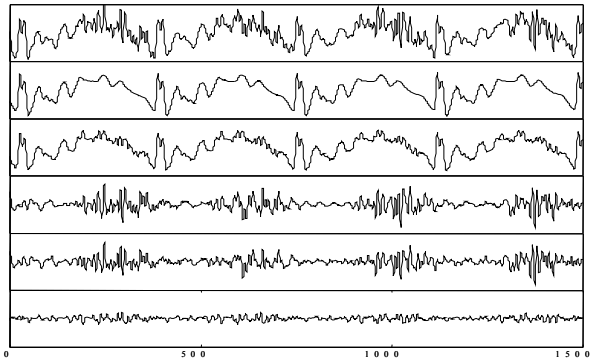


Figure 2 : Exemple de séparation H/B avec la méthode FTPS2 : voyelle /a/ avec modulation de la source de bruit et RHB = 0 dB. De haut en bas : signal mixte, signal H « vrai » et estimé, signal B « vrai » et estimé, différence entre signal « vrai » et estimé. L'axe des Y est arbitraire mais homogène entre les différentes figures. On obtient ici RSE = 9 dB.

4.2. Tests d'écoute informels

L'écoute des signaux a confirmé les bonnes performances de la méthode FTPS2, et l'amélioration par rapport à la méthode PSHF. Pour des RHB de 0 à 30 dB, le signal harmonique estimé avec la méthode FTPS2 n'est généralement pas distinguable du « vrai » signal harmonique, alors qu'il reste généralement un résidu de bruit significatif dans le signal harmonique estimé avec la méthode PSHF. La méthode PSHF semble bien souffrir du fait que les valeurs échantillonnées tous les quatre canaux de TFCT ne correspondent pas forcément exactement aux pics harmoniques si le signal est non-stationnaire. Pour des RHB faibles, les signaux séparés sont généralement de moins bonne qualité, c'est-à-dire moins proches des « vrais » signaux harmonique et de bruit, avec des qualités différentes pour les deux méthodes. Tous les signaux testés (« vrais » H et B, mixtes et séparés) sont disponibles online : www.icp.inpg.fr/~girin/HNS/HNS_demo.zip. Le lecteur est invité à se faire son propre jugement.

5. DISCUSSION

Bien qu'encourageants, les résultats précédents doivent être considérés prudemment, à cause de la dépendance de la méthode sur la précision des *pitch-marks* déjà mentionnée. En particulier, on peut s'attendre à ce que les mesures de phase des harmoniques de rang élevé soient significativement perturbées par les imprécisions du *pitch-marking*, puisque la variation de phase est égale à l'intégration temporelle de la fréquence. Cependant, ces limites doivent être fortement pondérées par deux points qui constituent le noyau de nos travaux actuels :

- D'abord, le filtrage passe-bas des paramètres spectraux pourrait intrinsèquement compenser ce bruit de mesure additif. En d'autres termes, le filtrage pourrait éliminer à la fois le bruit dû aux composantes bruitées de la parole et le bruit dû aux imprécisions de l'analyse. Ceci est vrai tant que la somme des contributions de ces bruits n'empêchent pas la trajectoire de fond des paramètres de phase d'émerger. Une étude plus poussée est nécessaire pour clarifier ce point. En particulier, l'influence de l'estimation automatique des *pitch-marks* doit être analysée, ainsi que les interactions entre les deux sources de bruit (mesures et parole elle-même).
- Ensuite, les *phases relatives* considérées dans cette étude peuvent être remplacées par les *phases absolues*, c'est-à-dire les valeurs de phase résultant de l'intégration temporelle des fréquences. En effet, à l'inverse des trajectoires de phase relative, les trajectoires de phase absolue peuvent être reconstruite à partir de mesures effectuées à des instants arbitraires. A l'inverse de la procédure de régularisation de la Section 2.2, cette reconstruction des trajectoires de phase absolue nécessite une procédure duale de dépliement [11] assez simple à implanter. Ainsi, l'estimation d'une trajectoire de phase absolue lisse à partir de mesures bruitées doit pouvoir conduire à un résultat au moins équivalent, avec l'avantage déterminant de ne pas dépendre des instants de mesure, comme les *pitch-marks* dans l'étude présente (par contre, la taille de la fenêtre d'analyse doit rester de l'ordre d'une période de signal pour capturer finement son évolution). Notons que le lissage des trajectoires de phase absolues peut aussi être obtenu par un filtrage passe-bas, mais aussi par des techniques alternatives comme la modélisation à long terme proposée dans [12].

6. REFERENCES

1. Stevens, K. N. *Acoustic phonetics*, MIT Press, Cambridge, MA, 1998.
2. Murphy, P. Perturbation-free measurements of the harmonics-to-noise ratio in voice signals using pitch-synchronous harmonic analysis, *J. Acoust. Soc. Am.*, 105(5), 2866-2880, 1999.
3. Stylianou, Y. Decomposition of speech signals into a deterministic and a stochastic part, *Proc. Int. Conf. Spoken Language Proc.*, Philadelphia, 1996.
4. Jackson, P. & Shadle, C. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech, *IEEE Trans. Speech Audio Proc.*, 9(7), 713-726, 2001.
5. Yegnanarayana, B., d'Alessandro, C. & Darsinos, V. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, 6(1), 1-11, 1998.
6. Serra, X. & Smith, J. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic + stochastic decomposition, *Comp. Music J.*, 14(4), 12-24, 1990.
7. Hermes, D. J. Synthesis of breathy vowels: Some research methods, *Speech Communication*, 10, 497-502, 1991.
8. Kang, G. & Everett, S. Improvement of the excitation source in the narrow-band linear prediction vocoder, *IEEE Trans. Acoust. Speech Sig. Proc.*, 33(2), 377-386, 1985.
9. George, E. & Smith, M, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.,* **5**(5), 389-406, 1997.
10. Rosenberg, A. E., Effect of glottal pulse shape on the quality of natural vowels, *J. Acoust. Soc. Am*. 49(2), 583-590, 1971.
11. R. J. McAulay & T. F. Quatieri, Speech analysis/ synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, 34(4), 744-754, 1986.
12. L. Girin, M. Firouzmand & S. Marchand, "Long term modeling of phase trajectories within the speech sinusoidal model framework," *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, 2004.

Adjonction de contraintes visuelles pour l'inversion acoustique-articulatoire

Blaise Potard, Yves Laprie

LORIA / Équipe PAROLE
Campus Scientifique - BP 239
54506 VANDŒUVRE-lès-NANCY CEDEX, France
Mél : {Blaise.Potard, Yves.Laprie}@loria.fr
http ://www.loria.fr/equipes/parole/

ABSTRACT

The goal of this work is to investigate audiovisual-to-articulatory inversion. It is well established that acoustic-to-articulatory inversion is an under-determined problem. On the other hand, there is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues : the view of visible articulators, i.e. jaw and lips, improves speech intelligibility. It is thus interesting to add constraints provided by the direct visual observation of the speaker's face. Visible data were obtained by stereo-vision and enable the 3D recovery of jaw and lip movements. These data were processed to fit the nature of parameters of Maeda's articulatory model. Inversion experiments show that constraints on visible articulatory parameters enable relevant articulatory trajectories to be recovered and substantially reduce time required to explore the articulatory codebook.

1. INTRODUCTION

La principale difficulté de l'inversion acoustico-articulatoire est le fait qu'il n'existe pas de relation directe du domaine acoustique vers l'articulatoire : un grand nombre de formes différentes de conduit vocal peuvent produire le même spectre de parole. C'est un problème sous-déterminé, puisqu'il y a plus d'inconnues que de données en entrée. Un des enjeux principaux réside dans l'étude de contraintes qui soient suffisamment restrictives et pertinentes d'un point de vue phonétique, de façon à éliminer des solutions manifestement peu réalistes.

La parole est un signal bimodal, comportant une composante acoustique, et une composante visuelle : la vue du locuteur. Ces deux modalités sont fortement corrélées et redondantes. Il a été observé à de nombreuses reprises que les locuteurs et auditeurs humains exploitent la nature multimodale de la parole, et plus particulièrement les indices articulatoires : l'intelligibilité de la parole augmente dans des conditions difficiles (déficiences auditives, environnement bruyant...) lorsque l'auditeur voit le locuteur[10, 1, 9, 4].

L'objectif du présent travail est d'adjoindre aux données acoustiques classiquement utilisées dans l'inversion acoustique-articulatoire, des données issues de l'observation des articulateurs visibles (lèvres et mâchoire), obtenues par stéréovision.

2. MÉTHODE D'INVERSION

Notre méthode d'inversion est fondée, comme beaucoup d'autres, sur l'analyse par synthèse, et le processus d'inversion comporte trois étapes.

Une étape préalable au processus d'inversion est la construction d'une table articulatoire, ou codebook, qui associe des vecteurs articulatoires (à 7 dimensions, correspondant aux 7 paramètres du modèle de Maeda) à leurs correspondants acoustiques, dans notre cas, le triplet des fréquences des 3 premiers formants. La force de notre méthode d'inversion réside dans la résolution acoustique quasi uniforme du codebook. Cette propriété est garantie par la façon dont est construite la table : on explore l'espace articulatoire récursivement en évaluant à chaque étape la linéarité locale de la relation articulatoire-acoustique [7]. Cette table est organisée de manière à retrouver facilement tous les vecteurs articulatoires qui permettent de générer un triplet de formants donné et est propre à chaque locuteur : sa construction nécessite une adaptation préalable du modèle articulatoire.

La première étape du processus d'inversion proprement dit consiste à générer un grand nombre de solutions potentielles à partir du codebook. Comme il existe *a priori* une infinité de vecteurs articulatoires permettant d'obtenir un vecteur acoustique il est nécessaire d'échantillonner l'espace des solutions de façon suffisamment concise mais précise pour trouver des solutions proches de la solution réelle.

La deuxième étape de notre méthode consiste à reconstruire une trajectoire articulatoire qui soit suffisamment régulière au cours du temps. Nous utilisons pour cela un algorithme de programmation dynamique qui minimise une fonction de coût représentant la « distance » couverte par les articulateurs.

La dernière étape consiste à améliorer la fidélité acoustique et la régularité articulatoire de la solution obtenue à l'étape précédente en utilisant un algorithme de régularisation variationnelle.

2.1. Exploration de l'espace nul de la relation articulatoire acoustique

Pour chaque vecteur acoustique représentée par les trois premières fréquences formantiques, le processus d'inversion consiste en la recherche de tous les hypercubes qui peuvent générer le triplet de formants observé. Il faut ensuite trouver un ensemble de solutions dans chacun de

ces cubes. Comme l'inversion consiste à trouver 7 paramètres à partir de 3, l'espace des solutions a *a priori* 4 degrés de liberté. La relation articulatoire acoustique (notée R) est supposée être localement linéaire au niveau du centre P_0 de l'hypercube (c'est-à-dire que l'application $P - P_0 \mapsto R(P) - R(P_0)$ est supposée être une application linéaire). Trouver l'ensemble des solutions n'est pas un problème trivial car il s'agit de trouver l'intersection d'un espace à 4 dimensions (l'espace nul de la relation précédente, c'est-à-dire l'ensemble des antécédents de 0 pour l'application linéaire) et d'un hypercube à 7 dimensions, ce que l'on ne sait pas faire de manière formelle. Une première approximation de l'intersection est obtenue par programmation linéaire. Puis l'espace nul est échantillonné, et l'appartenance à l'intersection de chacun des points est testée[8].

3. CONTRAINTES VISUELLES

3.1. Acquisition des données

Pour acquérir les données sur les articulateurs visibles nous avons utilisé un système d'acquisition et de suivi de données tridimensionnelles. Ce système a été réalisé par l'équipe de vision par ordinateur de notre laboratoire [11]. L'un de ses intérêts est d'être peu cher et facilement utilisable. Par ailleurs, il est plus flexible que les systèmes de *motion-capture* qui utilisent généralement des caméras infrarouges et des marqueurs collés sur la peau.

Il utilise simplement deux caméras, un PC, et des marqueurs peints qui ne perturbent pas l'articulation ; il permet une acquisition suffisamment rapide pour reconstituer de façon précise les trajectoires des points 3D.

Pour faire une reconstitution des mouvements des articulateurs en stéréovision, il est nécessaire d'être capable de suivre les même points physiques au cours du temps. Comme la peau naturelle n'est pas assez contrastée, nous avons choisi de peindre des marqueurs sur le visage du locuteur. Cette méthode permet de contrôler la taille, la densité et la position des points intéressants. Par exemple, nous avons peints 210 marqueurs sur le visage (46 sur les lèvres) pour permettre d'obtenir une information précise sur la déformation de la forme des lèvres (fig. 1), dans l'optique de construire une tête parlante de bonne qualité.

Dans le cas du corpus utilisé dans cette étude, élaboré dans l'optique d'étudier la variabilité interlocuteurs de la coarticulation labiale, nous avons choisi de peindre seulement 15 marqueurs sur le visage du locuteur (seulement 4 marqueurs sur les lèvres, fig. 2) de façon à conserver un temps de préparation raisonnable pour les sujets de l'étude. En plus des marqueurs utilisés pour étudier les mouvements des lèvres, nous avons placés 6 marqueurs dans la partie supérieure du visage de façon à compenser le mouvement global de la tête. Deux caméras monochromes sont utilisées, car leur vitesse d'acquisition (environ 120 images par seconde) étant plus rapide que celle des caméras couleurs, elles permettent de suivre des mouvements très rapides des articulateurs, par exemple les occlusives.

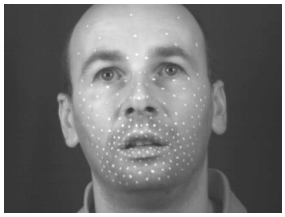


FIG. 1: 210 marqueurs blancs sont peints sur le visage du locuteur.

FIG. 2: Images en stéreovision de deux locuteurs, 15 marqueurs sont peints sur le visage de chaque locuteur.

3.2. Intégrer les données visuelles au modèle de Maeda

Le modèle articulatoire de Maeda[5] a été établi à partir d'images radiographiques de coupes sagittales du conduit vocal en y appliquant une analyse factorielle permettant le choix explicite des composantes linéaires. Les mouvements de la mâchoire, en particulier, peuvent être facilement déterminés en mesurant la position des incisives qui apparaissent très clairement sur les images. L'ouverture latérale des lèvres ne peut par contre pas être évaluée à partir des radiographies, et ce n'est donc pas un des paramètres du modèle.

Les données 3D du visage du locuteur permettent de mesurer directement l'étirement et l'ouverture des lèvres à partir de la position des marqueurs sur les lèvres (voir Fig. 2). La protrusion peut aussi être estimée à partir de ces points, mais comme il s'agit d'un mouvement complexe qui implique un « dépliement » des lèvres, les mouvements de marqueurs peints sur les lèvres dans le plan sagittal ne peuvent rendre compte que partiellement de ce mouvement complexe. Par conséquent, la protrusion est probablement légèrement sous-estimée.

Contrairement aux images radiographiques, les données 3D du visage du locuteur ne permettent pas de mesure précise des mouvements de la mâchoire. En effet, le mouvement des marqueurs peints sur le menton (que nous utilisons pour évaluer les mouvements de la mâchoire) est lié à la mâchoire, mais aussi à celui de la lèvre inférieure qui déplace ces marqueurs quand elle bouge. Par conséquent, le mouvement de la mâchoire n'est pas non plus connu avec précision.

3.3. Ajustement des données visuelles

À partir des données visuelles acquises, nous calculons 4 paramètres : l'ouverture de la bouche, l'étirement des lèvres et les mouvements de la mâchoire, ces paramètres se calculant facilement à partir de la position des marqueurs, et la protrusion des lèvres, dont l'évaluation est plus complexe.

- L'ouverture de la bouche est donnée par la distance entre les deux points des lèvres situés dans le plan sagittal.
- L'étirement des lèvres est la distance entre les deux points situés aux commissures des lèvres.
- Le mouvement de la mâchoire est la distance entre les points du menton et un point fixe. Nous prenons la moyenne des positions des 4 points du menton. En faisant cela, nous négligeons de façon implicite l'influence des mouvements des lèvres sur la position de ces points.
- La protrusion des lèvres est plus complexe à calculer. Le paramètre est déterminé en utilisant la distance des points des lèvres inférieure et supérieure à un plan de référence défini par la position moyenne des 4 points des lèvres.

Normalisation Comme notre objectif est d'utiliser les données 3D obtenues avec le système de stéréovision comme des contraintes sur les paramètres régissant les articulateurs visibles du modèle de Maeda, nous devons établir une correspondance entre les paramètres articulatoires du modèle et les paramètres observés que nous venons de définir. Les données géométriques mesurées par Maeda étaient centrées et normalisées avant d'être traitées par analyse factorielle. Chacun des 7 paramètres du modèle articulatoire de Maeda peut ainsi varier dans un intervalle de $\pm 3\sigma$ (où σ est l'écart-type du paramètre). Nous appliquons la même transformation aux paramètres issus des données tridimensionnelles : chacun des paramètres est centré autour de sa position moyenne et réduit.

Décorrélation L'étape de normalisation précédente permet d'obtenir des paramètres observés qui ont les mêmes dimensions que les paramètres articulatoires. Cependant, l'analyse factorielle de Maeda permettait en plus de cela de retirer l'effet de la mâchoire des autres paramètres de façon à obtenir des paramètres indépendants. Nous devons donc retirer l'effet des mouvements de la mâchoire des autres paramètres. De la même façon que Maeda nous calculons la corrélation entre la mâchoire et chacun des deux autres paramètres (l'ouverture et la protrusion des lèvres puisque nous n'utilisons pas l'étirement qui ne peut pas être utilisé dans le modèle) et soustrayons la corrélation des mesures normalisées.

Le principal problème de cette méthode est que, contrairement aux radiographies où le mouvement de la mâchoire est mesurable avec précision, le mouvement de la mâchoire inférieure n'est ici connu que de manière approchée. Cette étape de décorrélation ne permet d'obtenir par conséquent que des approximations de chacun des paramètres.

3.4. Intégration au processus d'inversion

Nous obtenons ainsi trois paramètres compatibles avec le modèle de Maeda, mais malheureusement imprécis. Nous devons donc compenser cette imprécision, ce que nous faisons en permettant aux paramètres visuels des solutions de l'inversion de varier dans un domaine assez important.

Pour cela, nous n'utilisons les paramètres observés que lors de la sélection des hypercubes : nous n'échantillonnons que les hypercubes dont les centres ont des paramètres visuels proches des paramètres observées. De cette manière l'imprécision des données visuelles ne se répercute pas directement sur les résultats de l'inversion et, comme nous le verrons plus tard, cette utilisation très simple permet d'accélérer considérablement le processus d'inversion et d'améliorer le réalisme des solutions.

Les paramètres observés ne sont donc pour l'instant utilisés que dans la première étape du processus d'inversion : la génération d'un grand nombre de solutions possibles. Dans la deuxième étape, on construit une trajectoire initiale optimale parmi cet ensemble de solutions en minimisant un critère biomécanique ou de régularité sur les paramètres articulatoires. Dans cette étude, on minimise un critère portant sur la « vitesse globale » des articulateurs à l'aide de l'algorithme de Ney[6]. Dans la troisième étape, on lisse cette trajectoire en utilisant un algorithme de régularisation variationnelle[3] pour améliorer la fidélité acoustique et la régularité de la solution.

4. EXPÉRIENCES

4.1. Corpus

La phrase que nous présentons dans cette étude est extraite d'un corpus de données 3D enregistré par 10 locuteurs français natifs (5 hommes et 5 femmes), chacun s'exprimant pendant environ 120 secondes. Ce corpus était principalement destiné à étudier la variabilité interlocuteurs de la coarticulation labiale, et comporte essentiellement des logatomes. Il comporte aussi une phrase « Le joaillier a broyé les cailloux de la voyageuse.» construite spécialement pour faciliter l'inversion, puisque la plupart des sons la composant sont des voyelles, des semi-voyelles ou d'autres sons voisés.

Nous présentons ici les résultats obtenus pour l'un des locuteurs. Le modèle articulatoire a été adapté au locuteur en utilisant la méthode de Galvan-Rodrigez[2]. Bien que nous ayons choisi une précision acoustique assez faible pour la construction du codebook (1 Bark), la précision acoustique moyenne reste très bonne : l'erreur RMS moyenne est d'environ 15 Hz pour F3.

4.2. Inversion de séquences

Nous présentons ici les résultats détaillés pour deux parties voisées de la phrase précédente : « joaillier » (/ʒɔɑje/) et « cailloux » (/kaju/) qui présentent des mouvements articulatoires intéressants. Nous ne présentons ici que les résultats non lissés, qui sont plus significatifs des forces et des faiblesses du système actuel.

La figure 3 présente les résultats de l'inversion pour la séquence « joaillier ». Nous affichons les trajectoires trouvées pour les 4 principaux paramètres (mâchoire, ouverture des lèvres, protrusion, position de la langue). Nous affichons aussi en trait pointillé les paramètres observés donnés comme contrainte. Comme on peut le voir sur le graphique de la mâchoire, l'inversion rencontre des difficultés au milieu de la séquence (instant 21400), pour la

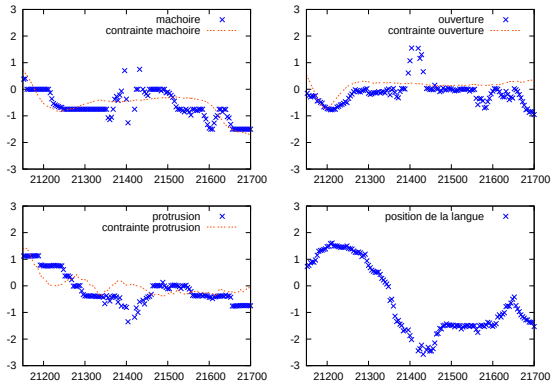


FIG. 3: Résultat de l'inversion pour « joaillier ». Pour chaque graphique, la courbe en trait fin pointillé correspond au paramètre issu des données visuelles, la courbe discontinue à la trajectoire du paramètre correspondant trouvée par le processus d'inversion. L'abscisse représente le temps (en ms), l'ordonnée la valeur du paramètre qui peut varier entre -3 et 3.

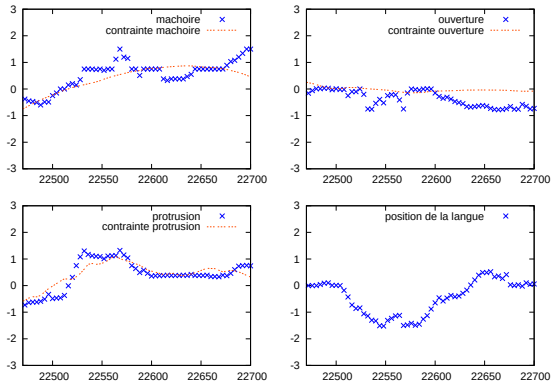


FIG. 4: Résultats de l'inversion pour « cailloux ».

transition /ɑj/. On retrouve la même irrégularité, moins marquée, pour les deux autres paramètres. Nous pouvons également observer que la position de la langue a une trajectoire cohérente avec le mouvement attendu : postérieure pour prononcer /ɔɑ/, elle avance beaucoup pour le /j/, puis recule légèrement pour /e/ (plus la valeur de ce paramètre est élevée, plus la langue est en arrière). Faute de place, nous n'affichons pas les trajectoires des autres articulateurs, d'autant que les évolutions de ces paramètres sont moins facilement interprétables.

La séquence « cailloux » (/aju/ en fait, notre système ne pouvant pas inverser le /k/) a été elle aussi inversée avec succès, et cela avec moins de difficulté pour satisfaire les contraintes visuelles. Comme cela apparaît sur la figure 4 les trois paramètres des articulateurs visibles ont des trajectoires très proches de leur contrainte. Le quatrième paramètre est la position de la langue, qui a, là aussi, une trajectoire phonétiquement réaliste bien que l'amplitude soit beaucoup plus faible (légèrement postérieure pour le /a/, antérieure pour /j/, postérieure pour /u/).

5. CONCLUSION ET PERSPECTIVES

Nous avons effectué le même travail sur les différents locuteurs de ce corpus avec des résultats très similaires. Cette approche du couplage entre paramètres visuels et acoustiques pour l'inversion audiovisuelle articulatoire est donc prometteuse. En effet, le modèle parvient à trouver des solutions satisfaisant les contraintes visuelles tout en prenant en compte les données acoustiques alors que les paramètres articulatoires visuels ne peuvent pas être récupérés avec une très grande précision. Par ailleurs, et il s'agit là d'un point essentiel, les trajectoires articulatoires récupérées sont réalistes d'un point de vue phonétique.

La poursuite de ce travail s'effectue suivant plusieurs axes. Nous travaillons à présent sur un autre corpus beaucoup plus précis géométriquement et plus long car il ne porte que sur une locutrice. Nous étudions en particulier une méthode pour établir une correspondance directe (plutôt que statistique) entre les données visuelles et les paramètres correspondants du modèle de Maeda. En effet, les paramètres obtenus actuellement ne sont pas tout à fait équivalents à ceux du modèle. Ensuite, nous souhaitons évaluer l'influence de l'adéquation géométrique du modèle articulatoire d'analyse au locuteur sur les trajectoires articulatoires inversées. Il est en effet probable que l'adjonction des contraintes conduise à des effets de compensation articulatoire artificiels.

RÉFÉRENCES

[1] C. Benoît, T. Mohamadi, and S. Kandel. Effect of phonetic context on audio-visual intelligibility of french. *Journal of Speech,Language and Hearing Research*, 37 :1195–1203, October 1994.

[2] A. Galván-Rdz. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives*. Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[3] Y. Laprie and B. Mathieu. A variational approach for estimating vocal tract shapes from the speech signal. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 929–932, Seattle, USA, May 1998.

[4] B. Le Goff. Automatic modeling of coarticulation in text-to-visual speech synthesis. In *Eurospeech'97 Proceedings*, volume 3, pages 1667–1670, Rhodes, Greece, 1997. European Speech Communication Association.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[6] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2) :163–173, March 1983.

[7] Slim Ouni and Yves Laprie. Improving acoustic-to-articulatory inversion by using hypercube codebooks. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, China*, volume II, pages 178–181, October 2000.

[8] Slim Ouni and Yves Laprie. Studying articulatory effects through hypercube sampling of the articulatory space. In *17th International Congress on Acoustics , Rome, Italy*, volume 4, September 2001.

[9] J. Robert-Ribes, J-L. Schwartz, and P. Escudier. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, 9 :323–346, 1994.

[10] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *JASA*, 26(2) :212–215, 1954.

[11] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low cost stereovision based system for acquisition of visible articulatory data. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, pages 145–150, Vancouver, 2005.

Estimation des instants de fermeture basée sur un coût d'adéquation du modèle LF à la source glottique

Damien Vincent [(1)], *Olivier Rosec* [(1)], *Thierry Chonavel* [(2)]

[(1)] France Telecom, Division R&D
2, Avenue Pierre Marzin - 22307 Lannion
{damien.vincent,olivier.rosec}@francetelecom.com
[(2)] École Nationale Supérieure des Télécommunications de Bretagne, Département Signal et Communication,
Technopôle Brest-Iroise, CS 83818, 29285 Brest Cedex, France.
thierry.chonavel@enst-bretagne.fr

ABSTRACT

An algorithm for GCI (Glottal Closure Instants) estimation is presented in this paper. It relies on a source-filter model of speech production using a LF model for the source component. From this source-filter decomposition, a ratio which measures the goodness of fit of the LF source model is introduced in the GCI estimation procedure together with fundamental frequency constraints. Then, a Viterbi algorithm is applied to extract the most likely GCI sequence. Experiments performed on a real speech database show that the proposed method outperforms existing approaches.

1. INTRODUCTION

L'estimation des instants de fermeture de glotte est un problème récurrent en traitement de la parole. Ces instants sont cruciaux en analyse du signal de parole, car leur localisation précise est nécessaire pour estimer le signal de source glottique ainsi que pour caractériser la qualité vocale. Une autre application liée à la synthèse vocale par concaténation concerne le marquage pitch-synchrone des bases de données acoustiques, opération nécessaire pour la mise en oeuvre d'algorithmes de modification prosodique tel que TD-PSOLA [6]. Pour la détermination de ces GCI, plusieurs méthodes ont été proposées telles que celles basées sur la fonction de retard de groupe [7] qui exploitent des propriétés basiques des signaux à phase minimale, ou encore celles reposant sur des algorithmes de programmation dynamiques pour estimer une séquence de GCI en accord avec une mesure préalable de la fréquence fondamentale F_0 [5].

Cependant, les approches existantes ne tiennent pas compte de façon explicite de la structure du signal glottique. Dans cet article, nous utilisons une mesure obtenue à partir d'une décomposition source-filtre du signal de parole afin de localiser les GCI potentiels et donc de mieux contraindre le problème d'estimation. Nous définissons alors une fonction de coût combinant cette information et une mesure de F_0 préalablement obtenue. L'estimation de la séquence de GCI est alors obtenue par minimisation de cette séquence via un algorithme de programmation dynamique. Le papier est organisé comme suit. En section 2, nous présentons le modèle source-filtre utilisé et définissons une mesure d'adéquation à ce modèle. La section 3 détaille l'algorithme d'estimation des GCI proprement dit et la section 4 décrit les expériences destinées à valider la méthode proposée.

2. MODÈLE ARX

De nombreux modèles de production de la parole font l'hypothèse que le signal de parole résulte du filtrage linéaire de l'excitation glottique par le conduit vocal. Dans une telle décomposition source-filtre, la partie source, appelée Dérivée de l'Onde de Débit Glottique (DODG), correspond au signal produit au niveau de la glotte après prise en compte de l'effet de radiation des lèvres, approximé par une dérivation. La partie filtre désigne, quant à elle, les résonances du conduit vocal.

Lors de la production de sons voisés, les cordes vocales entrent en vibration, ce qui se traduit par une DODG quasi-périodique. Plusieurs modèles ont été proposés pour modéliser la DODG ainsi produite. Nous considérons ici le modèle LF [4] qui permet une paramétrisation de la forme de la DODG à l'aide de trois paramètres. La figure 1 représente une onde obtenue par ce modèle.

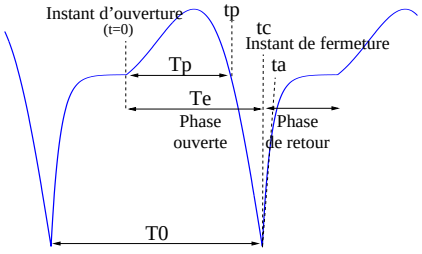


FIG. 1: Le modèle LF

Une composante stochastique est également présente pour modéliser les différents phénomènes aléatoires (irrégularité de la DODG, bruit de friction, etc...). Etant données ces hypothèses, un son $s(n)$ peut être modélisé par un processus ARX (Auto Regressive eXogenous) défini par l'équation suivante :

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + b_0 u(n) + e(n) \quad , \qquad (1)$$

dans laquelle les a_k sont les coefficients du modèle AR caractérisant le conduit vocal, $u(n)$ désigne la DODG dont l'amplitude est contrôlée par le coefficient b_0 et $e(n)$ est le résidu.

L'estimation des paramètres du modèle ARX est très complexe, car l'optimisation selon les paramètres du modèle LF est un problème non-linéaire. Cependant, lorsque la source LF est fixée, le filtre peut être estimé par des méthodes de moindres carrés classiques. Sur la base de ce

constat, nous utilisons ici une méthode d'estimation efficace proposée dans [8]. Cette méthode consiste dans un premier temps à effectuer une recherche exhaustive dans un espace de DODG quantifiées, puis à procéder à une optimisation locale.

3. PROCÉDURE D'ESTIMATION DES GCIS

3.1. Principe de la méthode

La méthode que nous présentons dans cette section vise à estimer la séquence de GCI optimale au sens d'un critère combinant deux types de métriques. D'une part, les GCI estimés doivent être en accord avec le modèle de source défini précédemment, ce qui suggère d'associer à chaque instant potentiel, un coût cible. D'autre part, un coût dit de concaténation doit contraindre deux GCI consécutifs à être proche d'une mesure locale de la période fondamentale supposée connue : en pratique, la fréquence fondamentale sera estimée à l'aide de l'algorithme YIN [3] et en utilisant une interpolation par des splines cubiques pour obtenir une estimée pour chaque échantillon. Formellement, il s'agit donc de déterminer la séquence de GCI minimisant la fonction de coût définie par :

$$C = \sum_{l=1}^{L} C_{\text{cible}}(t_c^l) + \sum_{l=2}^{L} C_{\text{concat}}(t_c^l, t_c^{l-1}) \quad , \qquad (2)$$

où t_c^l désigne le $l^{\text{ème}}$ GCI candidat et où C_{cible} et C_{concat} désignent respectivement le coût cible et le coût de concaténation à satisfaire. Notons que le nombre L d'instants de fermeture de glotte n'est *a priori* pas connu. La résolution d'un tel problème d'optimisation peut être obtenue par programmation dynamique. Les détails relevant de l'implémentation algorithmique seront présentés en section 3.4.

3.2. Coût cible

Le modèle ARX présenté en section 2 permet de représenter le signal de parole comme la convolution d'une approximation LF de la source glottique par le filtre modélisant le conduit vocal à laquelle se rajoute un terme d'erreur de modélisation. Cette prise en compte explicite de l'information *a priori* sur la source glottique confère au modèle ARX une meilleure capacité de modélisation du signal de parole. De ce fait, l'erreur quadratique moyenne du résidu issu du modèle ARX est toujours inférieure à celle obtenue par une modélisation AR.

Cette constatation suggère de définir pour tout instant candidat t_c^l une mesure d'adéquation au modèle de la forme :

$$C_{\text{cible}}(t_c^l) = \frac{E_{LF}(t_c^l)}{E_0(t_c^l)} \quad , \qquad (3)$$

où $E_{\text{LF}}(t_c^l)$ désigne l'erreur quadratique moyenne issue du modèle ARX en utilisant la source optimale dont l'instant de fermeture est situé en t_c^l et où $E_0(t_c^l)$ est l'erreur quadratique moyenne obtenue par prédiction linéaire. Cette mesure normalisée tend vers 0 lorsque le son analysé est purement voisé et qu'il suit parfaitement le modèle LF ; elle tend vers 1 si le signal est non voisé ou si le modèle LF est très éloigné du signal glottique réel.

3.3. Le coût de concaténation

Le coût de concaténation $C_{concat}(t_c^l, t_c^{l-1})$ entre le $l^{\text{ième}}$ GCI t_c^l et le précédent t_c^{l-1} vise à pénaliser des distances $\Delta t_c^l = t_c^l - t_c^{l-1}$ entre ces deux GCI trop éloignées de la période fondamentale estimée en t_c^l. Le coût induit doit cependant être en adéquation avec la confiance accordée à la période fondamentale estimée : lorsque cette confiance est faible, l'estimation des GCIs sera d'avantage basée sur le coût cible ; tandis que si la confiance est élevée et si le coût cible ne permet pas de discriminer les instants de fermeture, le coût de concaténation sera privilégié, le processus d'estimation s'apparente dans ce cas à un mécanisme d'interpolation du GCI courant à partir du précédent. La distance CMNDF (*Cumulative Mean Normalized Difference Function*) introduite dans [3] est ainsi utilisée comme mesure de confiance pour moduler le coût de concaténation. Cette distance est définie par :

$$d_n^2(\tau) = \begin{cases} 1 & \text{si } \tau = 0, \\ \frac{d_n(\tau)}{\sum_{k=1}^{\tau} d_n(k)} & \text{sinon,} \end{cases}$$

où $d_n(\tau) = \sum_{k=-K}^{K} \left(s(n+k) - s(n+k-\tau)\right)^2$ correspond à la fonction différence sur une fenêtre de longueur $2K+1$.

La modulation du coût de concaténation par cette mesure de confiance sera réalisée en deux étapes. Tout d'abord, la distance CMNDF sert à définir les périodes minimale et maximale entre deux GCI consécutifs, les fréquences associées étant obtenues à partir de la fréquence fondamentale estimée $f_0(t_c^l)$ par :

$$\begin{array}{l} \frac{f_0(t_c^l)}{f_0^{\min}} = \frac{f_0^{\max}}{f_0(t_c^l)} \\ \ln\left(\frac{f_0^{\max}}{f_0(t_c^l)}\right) = \gamma \frac{\min(d^2_{t_c^l}(T_0(t_c^l)); 1) + \delta}{1+\delta} \end{array} \quad . \qquad (4)$$

Le paramètre δ autorise une certaine variation de la période entre 2 GCI même si la distance CMNDF est nulle (c'est à dire correspondant à une confiance très élevée) tandis que le paramètre γ correspond à un facteur d'échelle : en pratique, $\delta = 0.15$ et $\gamma = 0.53$ ce qui donne $\frac{f_0^{\max}}{f_0(t_c^l)} = 1.07$ pour $d^2_{t_c^l}(T_0) = 0$ et $\frac{f_0^{\max}}{f_0(t_c^l)} = 1.70$ pour $d^2_{t_c^l}(T_0) \geq 1$. A partir de ces fréquences minimale et maximale, nous en déduisons un premier coût de concaténation représenté sur la figure 2 et donné par :

$$C^1_{\text{concat}}(t_c^l, t_c^{l-1}) = g\left(\frac{f_s}{t_c^l - t_c^{l-1}}\right) \quad . \qquad (5)$$

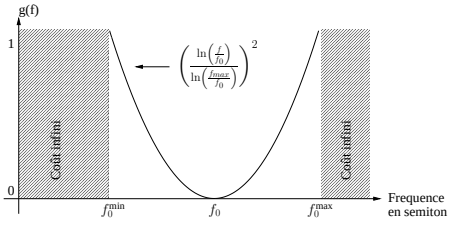


FIG. 2: Pénalité appliquée en fonction de l'écart par rapport à la fréquence fondamentale estimée.

Pour rendre l'estimation des GCI plus robuste aux erreurs d'estimation de la fréquence fondamentale, une seconde

modulation du coût de concaténation définie par

$$C^2_{\text{concat}}(t_c^l, t_c^{l-1}) = \min\left(d^2_{t_c^l}(\Delta t_c(l)) - \min_\tau d^2_{t_c^l}(\tau); 1\right) \quad (6)$$

est introduite de manière à favoriser des périodes Δt_c^l correspondant à des valeurs faibles de la fonction CMNDF. Nous obtenons au final la fonction de concaténation suivante :

$$C_{\text{concat}}(t_c^l, t_c^{l-1}) = C^1_{\text{concat}}(t_c^l, t_c^{l-1})\, C^2_{\text{concat}}(t_c^l, t_c^{l-1}) \quad (7)$$

La figure 3 illustre l'intérêt du coût C^2_{concat} : la période T_1 se trouve favorisée tout autant que la période estimée T_0.

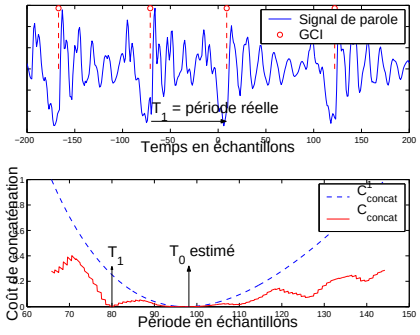


FIG. 3: Exemple de coût de concaténation sur une zone où les GCI du signal de parole sont irrégulièrement espacés

3.4. *Considérations pratiques*

Le nombre L de GCIs n'étant pas connu a priori, se pose le problème de la terminaison de l'algorithme. Un critère basé sur la minimisation de C (défini par l'équation 2) vis à vis de L n'est pas valable car les fonctions de coût cible et de concaténation sont toujours positives : le minimum de C par rapport à L serait donc obtenu en prenant $L = 1$ et en ne sélectionant que le GCI qui correspond au minimum global du coût cible. Le problème de terminaison peut se résoudre en considérant le problème d'estimation des GCI comme un problème de plus court chemin : les premier et dernier GCI sont d'abord contraints à être respectivement dans un intervalle de départ $[t_1^s, t_2^s]$ et d'arrivée $[t_1^e, t_2^e]$, l'algorithme d'estimation doit ensuite déterminer le plus court chemin entre ces 2 intervalles en plaçant L GCI le long du chemin optimal.

L'espace d'état associé à l'algorithme de programmation dynamique est composé de $N = t_2^e - t_1^s$ échantillons. La complexité de chaque itération n'est cependant pas $O(N^2)$ car pour un état courant donné, un grand nombre d'états précédents est interdit : la complexité est réduite à $O(N(T_0^{max} - T_0^{min}))$ où T_0^{max} et T_0^{min} sont les périodes associées à f_0^{min} et f_0^{max} définis par l'équation 4. Pour réduire d'avantage la complexité, l'algorithme pourra être appliqué sur chaque partie voisée du signal séparément ce qui permet de réduire le nombre d'états.

4. TESTS

L'évaluation a été réalisée sur la base *arctic* [1] qui fournit à la fois les signaux de parole et les enregistrements électroglottographiques (EGG) correspondants. La dérivee du signal EGG (DEGG) permet d'extraire facilement les instants de fermeture réels car ceux-ci correspondent à des pics très marqués sur ce signal. Ces GCI de référence sont nécessaires pour évaluer de manière objective les performances de notre algorithme.

4.1. *Détermination des GCI de référence*

L'évaluation étant réalisée sur une base entière de signaux de parole, une méthode manuelle d'extraction des pics correspondant aux GCI n'est pas réaliste. L'algorithme 1 permet de déterminer automatiquement les GCI, les GCIs étant extraits selon une confiance décroissante : la confiance est mesurée directement par l'amplitude du pic correspondant sur le signal DEGG. Il est à noter que l'algorithme utilise une estimée de la période fondamentale T_0 ce qui pourrait le rendre dépendant des performances de l'estimateur de T_0 ; l'algorithme s'avère cependant robuste aux erreurs d'estimation de T_0.

Algorithme 1 : Extraction de l'ensemble G des GCI de référence à partir du signal DEGG $y(t)$

$\forall t : c(t) = 1$
Tant que $M = \max_t(c(t)y(t)) \geq$ *seuil* **faire**
  $t_c \leftarrow \text{argmax}_t(c(t)y(t))$
  $G \leftarrow G \bigcup \{t_c\}$
  $t_1 \leftarrow \max\left\{t \leq t_c - 1\text{ms} / \left(1 - \frac{t_c - t}{T_0(t_c)}\right) M \leq y(t)\right\}$
  $t_2 \leftarrow \min\left\{t \geq t_c + 1\text{ms} / \left(1 - \frac{t - t_c}{T_0(t_c)}\right) M \leq y(t)\right\}$
  $c([t_1; t_2]) = 0$

4.2. *Critères de performance*

A l'aide de l'algorithme 2, les quatre ensembles suivants sont construits : l'ensemble des indices de référence non détectés (ND), l'ensemble des GCI estimés qui sont des fausses alarmes (FA), l'ensemble contenant les appariements de GCI de référence et estimé qui correspondent à des erreurs supérieures à 2.5ms (erreurs grossières EG) et enfin l'ensemble E correspondant aux couples de GCI de référence et estimés dont l'erreur d'estimation est inférieure à 2.5ms. A partir de ces ensembles sont obtenus les mesures de performance suivantes : le taux de non détections TND $= \frac{\text{card}\{ND\}}{N_r}$ où N_r correspond au nombre de GCI de référence, le taux de fausses alarmes TFA $= \frac{\text{card}\{FA\}}{N_r}$, le taux d'erreurs grossières TEG $= \frac{\text{card}\{EG\}}{N_r}$ et la variance sur E de l'erreur d'estimation des GCI.

Algorithme 2 : Association entre les GCI de référence et les GCI estimés

K ensemble des indices des GCI de référence
L ensemble des indices des GCI estimés
Tant que $K \neq \emptyset$ *et* $L \neq \emptyset$ **faire**
  $(k_m, l_m) = \text{argmin}_{(k,l) \in K \times L}\, t_c(k) - \hat{t}_c(l)$
  $\Delta = t_c(k_m) - \hat{t}_c(l_m)$
  $K \leftarrow K \backslash \{k_m\}$ et $L \leftarrow L \backslash \{l_m\}$
  Si $\Delta >$ 5ms : ND = ND $\bigcup k_m$ et FA = FA $\bigcup l_m$
  Si $\Delta \in [2.5; 5\text{ms}]$: EG = EG $\bigcup (t_c(k_m), \hat{t}_c(l_m))$
  Si $\Delta < 2.5ms : E = E \bigcup (t_c(k_m), \hat{t}_c(l_m))$
ND = ND $\bigcup K$
FA = FA $\bigcup L$

Deux types d'évaluation sont réalisés : l'une en utilisant l'ensemble des GCI $t_c(l)$ de référence, l'autre en n'utilisant que les GCI qui sont régulièrement espacés. Le GCI l est dit irrégulier si l'une des 2 conditions suivantes est réalisées : $\frac{t_c(l)-t_c(l-1)}{(t_c(l+2)-t_c(l-2))/4} \notin [0.8; 1.2]$ ou $\frac{t_c(l+1)-t_c(l)}{(t_c(l+2)-t_c(l-2))/4} \notin [0.8; 1.2]$; c'est à dire si la période à gauche ou à droite dévie trop de la période moyenne.

4.3. Résultats

Les performances de l'algorithme proposé sont comparées à celles de l'algorithme DYPSA [5]. L'estimation des GCI par DYPSA est basée sur l'utilisation des délais de groupe pour déterminer une liste d'instants de fermeture candidats : la fonction de délai de groupe EW (*Energy Weighed Group Delay*) définie dans [2] est appliquée au résidu LPC en utilisant une longueur de fenêtre de 10ms pour les voix d'homme et de 7ms pour les voix de femme. La séquence la plus probable est ensuite déterminée à l'aide d'un algorithme de programmation dynamique qui permet entre autres d'introduire des contraintes de régularité des périodes fondamentales obtenues à partir de la différence entre deux GCI consécutifs.

Le tableau 1 présente les écarts-types, les taux d'erreurs grossières, de non détections et de fausses alarmes pour l'algorithme proposé et la méthode DYPSA. Dans la configuration de test C1 comprenant les GCI irréguliers, l'algorithme DYPSA présente un taux de mauvaise détection plus élevé (3.42%). En supprimant les GCI irréguliers des statistiques, les performances des deux algorithmes sont bien meilleures, l'algorithme proposée présente cependant une variance d'estimation plus faible que DYPSA et les taux TEG, TFA et TND sont également plus faibles. La figure 4 montre sur un exemple que l'algorithme proposé est capable d'estimer correctement les GCI sur les zones stationnaires mais aussi sur des zones où les GCI sont irrégulièrement espacés. A notre sens, deux raisons peuvent expliquer les meilleures performances de notre algorithme. Tout d'abord, la contrainte de régularité des GCI est beaucoup plus souple dans l'algorithme proposé ce qui permet d'obtenir de bons résultats sur les zones où les GCI sont irrégulièrement espacés ; ensuite, le coût cible utilisé semble être plus discriminant que les fonctions de délai de groupe qui d'une part amènent à prendre en compte certains instants d'ouverture très prononcés comme GCI candidat, d'autre part peuvent aboutir à une mauvaise précision lorsque l'instant de fermeture est peu marqué.

Test	Algorithme	σ	TEG	TFA	TND
C1	Proposé	0.37	0.73	0.95	0.57
	DYPSA	0.38	1.20	1.35	3.42
C2	Proposé	0.25	0.09	0.09	0.08
	DYPSA	0.30	0.35	0.25	0.90

TAB. 1: Comparaison des deux méthodes : variance d'estimation (en ms), taux TEG, TFA et TND (en %) pour la configuration de test C1 (utilisant tous les GCI de références) et la configuration de test C2 (utilisant uniquement les GCIs régulièrement espacés).

5. CONCLUSION

La méthode proposée permet d'estimer les instants de fermeture avec une bonne précision, tout en gardant des taux de fausses alarmes et de non détections faibles. Ces bonnes performances sont imputables : i) au choix d'un coût cible en accord avec les mécanismes de production de la parole et ii) à l'utilisation de contraintes de continuité permettant un bon compromis entre le respect de la période fondamentale estimée et l'adéquation au coût cible. A partir de ces coûts, la séquence optimale de GCI peut être déterminée en utilisant un algorithme de plus court chemin appliqué sur l'ensemble des échantillons du signal. Des études complémentaires restent néanmoins nécessaires afin d'une part de caractériser plus précisément les erreurs produites par l'algorithme proposé et d'autre part de valider les résultats obtenus sur d'autres bases de parole.

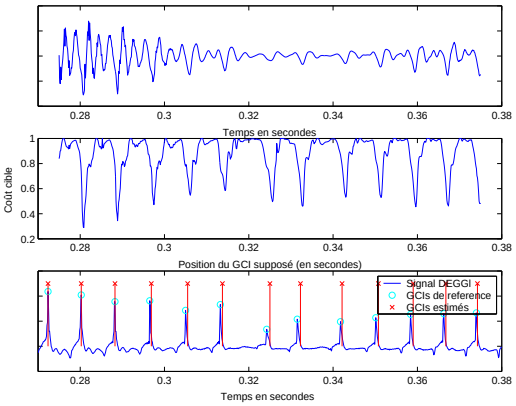


FIG. 4: Résultat de l'estimation des GCI sur un signal correspondant au mot anglais '*was*'. De haut en bas : (a) le signal de parole, (b) le coût cible $C_{\text{cible}}(t_c)$, (c) le signal DEGG et les GCI estimés.

RÉFÉRENCES

[1] Arctic speech database. http ://festvox.org/cmu_arctic/.

[2] M. Brookes, P.A. Naylor, and J. Gudnason. A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Trans. on Speech and Audio Processing*, 2006.

[3] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4) :1917–1930, 2002.

[4] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4 :1–13, 1985.

[5] A. Kounoudes, P.A. Naylor, and M. Brookes. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. *IEEE ICASSP*, May 2002.

[6] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 :453–467, 1990.

[7] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. on Speech and Audio Processing*, 3(5) :325–333, 1995.

[8] D. Vincent, O. Rosec, and T. Chonavel. Estimation of LF glottal source parameters based on ARX model. *Interspeech*, pages 333–336, 2005.

Codage à bas débit des paramètres LSF par quantification vectorielle codée par treillis

M. BOUZID, A. DJERADI, B. BOUDRAA

Laboratoire Communication Parlée et Traitement du signal
Equipe Codage de la Parole, Faculté d'Electronique et d'Informatique
Université *USTHB*, BP 32, El-Alia, Bab-Ezzouar, ALGER, 16111, ALGERIE
Fax de la faculté: 213-21- 24.76.07. Email: mbouzid@yahoo.com

ABSTRACT

Speech coders operating at low bit rates necessitate efficient encoding of the linear predictive coding (LPC) coefficients. Line spectral Frequencies (LSF) parameters are currently one of the most efficient choices of transmission parameters for the LPC coefficients. In this paper, an optimized trellis coded vector quantization (TCVQ) scheme for encoding the LSF parameters is developed. When the selection of a proper distortion measure is the most important issue in the design and operation of the encoder, an appropriate weighted distance measure has been used during the TCVQ construction process. Using this distance, we will show that our LSF TCVQ encoder performs better than the encoder conceived with the unweighted distance.

1. INTRODUCTION

Dans des applications de codage de la parole à bas débit, l'information spectrale à court terme d'un signal de parole est souvent modélisée par la réponse fréquentielle d'un filtre tous-pôles de fonction de transfert $H(z) = 1/A(z)$, avec $A(z) = 1 + a_1 z^{-1} + \ldots + a_{10} z^{-10}$. Les 10 coefficients $\{a_i\}_{i=1,2,\ldots,10}$ de ce filtre, connus sous le nom de coefficients de prédiction linéaire (LPC) [1], sont dérivés du signal d'entrée en utilisant une analyse par prédiction linéaire (LP) sur chaque trame du signal parole. Dans la pratique, on ne quantifie pas directement les coefficients LPC car ils ne sont pas appropriés au codage. Plusieurs transformations équivalentes ont été développées afin de les convertir en paramètres beaucoup plus appropriés à la quantification. Parmi les représentations qui se sont avérés efficaces, les fréquences de raies spectrales LSF (Line Spectral Frequencies) sont sans doute les plus utilisées [2]. Les paramètres LSF, qui sont liés aux zéros de polynômes dérivés de $A(z)$, présentent un certain nombre de propriétés intéressantes [1, 3]. Exploitant ces propriétés, divers schémas de codage basés sur la quantification scalaire et vectorielle ont été suggérés pour la quantification efficace des paramètres LSF. Les schémas de codage à base d'un quantificateur scalaire (SQ) sont intéressants dûs à leur niveau bas de complexité; cependant, ils accomplissent la qualité de quantification transparente à des débits hauts [4]. Un quantificateur vectoriel (VQ) peut réaliser la qualité de quantification transparente à des débits binaires plus bas [3]. Cependant, il est plus complexe et exige des tailles mémoire élevées. Pour réduire la complexité des calculs et les exigences en taille mémoire, divers schémas à base de VQs sous-optimaux, comme les VQs multi-étages [5], les VQs divisés (Split) [3],…, ont été proposés dans le passé pour coder les paramètres LSF.

Dans cet article, nous présentons un système de codage optimisé à quantification vectorielle codée par treillis TCVQ (Trellis Coded Vector Quantization) pour coder les paramètres LSF à bas débit. Dans ce système, que nous avons appelé LSF-OTCVQ, la dépendance intra-trame entre les plus proches paires de paramètres LSF sera exploitée en utilisant des dictionnaires bidimensionnels (2-D). Connaissant que le choix d'une mesure de distorsion appropriée est une question importante dans la conception d'un système VQ, nous avons utilisé une mesure de distance pondérée dans l'étape de conception et de fonctionnement de notre encodeur LSF-OTCVQ.

2. QUANTIFICATION VECTORIELLE CODEE PAR TREILLIS

La quantification scalaire codée par treillis (TCQ) est une forme améliorée du codage en treillis traditionnel. Elle attribue aux branches du treillis des sous-ensembles de niveaux de quantification plutôt que des niveaux de quantification individuels [6,7,8]. L'approche TCQ, qui a été motivée par la formulation d'Ungerboeck de la Modulation Codée par Treillis (TCM) [9], utilise un dictionnaire (alphabet) structuré avec un ensemble augmenté de niveaux de quantification.

Pour coder une source sans mémoire de dimension k=1 par un codeur TCQ opérant à un débit de R bits/ échantillon (bpe), on peut utiliser n'importe quel treillis d'Ungerboeck pour modulation d'amplitude [9]. L'alphabet augmenté (doublé) peut être choisi comme l'ensemble des niveaux de quantification obtenus par un SQ de Lloyd-Max de débit R+1 bpe. Ainsi, un alphabet se composant deux fois d'autant de grandeurs scalaires, à savoir 2^{kR+1} (k=1), est construit. Cependant, lors du processus de codage, la structure du treillis ramène le nombre augmenté de niveaux de quantification au débit de codage désiré (R). Par conséquent, seulement 2^{kR} de ces niveaux peuvent être utilisés pour représenter un échantillon de source à un instant donnée. Basé sur les règles de la partition d'ensemble d'Ungerboeck [9], ces

niveaux sont ensuite divisés en 4 sous-ensembles (D_0, D_1, D_2, D_3) et étiquetés aux branches du treillis. Pour quantifier la séquence source, l'algorithme de Viterbi est utilisé pour trouver le chemin optimal à travers le treillis. La séquence des niveaux choisis par cet algorithme est représentée par une séquence de bits indiquant le chemin optimal (séquence de sous-ensembles des niveaux de quantification) en plus d'une séquence de bits (mots-code) nécessaire pour indiquer les niveaux choisis à l'intérieur des sous-ensembles du chemin optimal. A la réception, le décodeur reconstruit la source quantifiée comme suit: la séquence de bits indiquant le chemin optimal à travers le treillis est utilisée comme entrée du codeur convolutionnel du TCQ; sa sortie sélectionne les sous-ensembles appropriés. Les mots-code de la seconde séquence sont utilisés pour indiquer les niveaux corrects à l'intérieur de chaque sous-ensemble choisi. Le système TCQ est complètement défini par la structure du treillis, les niveaux de quantification et la partition d'ensemble. Un exemple de codage d'une source uniformément distribuée sur l'intervalle [-A A] par un codeur TCQ à 4 états de débit R=2 bpe est illustré à la figure 1.

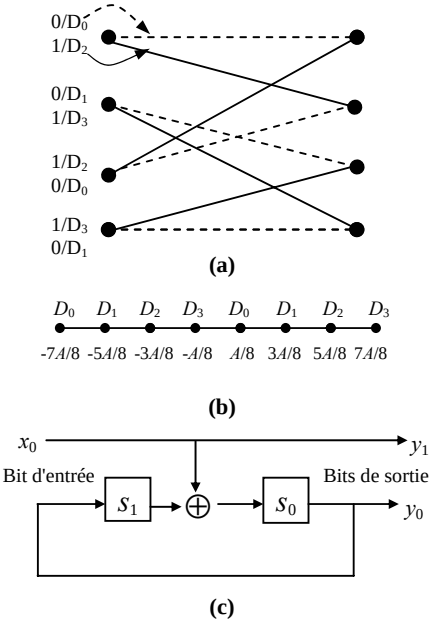


Figure 1 : Encodeur TCQ de débit R =2 bits/échantillon,
(a) Section du treillis étiqueté d'Ungerboeck à 4 états,
(b) Niveaux doublés de l'alphabet TCQ et la partition,
(c) Codeur convolutionnel TCQ correspondant.

Bien que les performances d'un codeur TCQ [6, 10] sont dans la plupart des cas proches de la limite théorique $D(R)$, des améliorations sont toujours possible en généralisant sa structure au cas vectoriel. On parle alors de la quantification vectorielle codée par treillis (TCVQ) [7, 8, 10]. La structure d'un codeur TCVQ est similaire à celle d'un système TCQ, avec une augmentation de la complexité des calculs qui est due à la recherche de vecteurs dans les sous-ensembles. Une meilleure conception du dictionnaire-TCVQ initial est réalisée par l'algorithme LBG d'un VQ (LBG-VQ) [11]. Une fois que les vecteurs-code sont déterminés, ils sont divisés dans des sous-ensembles et les sous-ensembles sont étiquetés aux branches du treillis suivant les même règles de la partition d'ensemble d'Ungerboeck [9]. La partition d'ensemble qui est plutôt simple dans le cas scalaire, n'est pas une tâche facile dans le cas vectoriel. Dans la conception de nos encodeurs TCVQ, nous avons utilisé l'heuristique décrite dans [8] pour réaliser la partition d'ensemble du dictionnaire étendu.

Des exemples de résultats de simulation, obtenus lors du codage TCVQ de sources sans mémoire Gaussiennes, sont listés dans table 1. Pour différents débits fractionnaires, les résultats sont donnés en termes de rapport signal sur bruit (RSB) en dB, avec les performances du VQ-LBG conventionnel et les fonctions Distorsion-Débit $D(R)$ correspondantes.

Table 1 : Performances RSB de la TCVQ (à des débits fractionnaires) d'une source Gaussienne.

Débit (bpe)	Dim. k	Nombre d'états							VQ LBG	$D(R)$
		4	8	16	32	64	128	256		
0.66	6	3.34	3.39	3.41	3.42	3.45	3.48	3.49	3.03	4.01
0.75	4	3.72	3.78	3.80	3.82	3.87	3.90	3.93	3.35	4.51
0.80	5	3.96	4.04	4.07	4.08	4.14	4.18	4.20	3.70	4.82

Pour un même débit de codage, ces résultats montrent que les performances de la TCVQ sont nettement supérieures à celles du VQ-LBG standard. Comparé aux limites $D(R)$, des améliorations sont toujours possibles.

Pour améliorer d'avantage les performances de la TCVQ, une procédure d'optimisation de la conception du dictionnaire-TCVQ a été élaborée. Cette procédure d'optimisation par entraînement utilise l'étape standard de l'algorithme LBG [11] appliquée à une séquence d'apprentissage de la source à coder. Elle consiste à mettre à jour le dictionnaire TCVQ en remplaçant chaque vecteur-code avec la moyenne de tous les vecteurs source attribués à ce vecteur. Ceci mène à un algorithme itératif de conception pour l'encodeur TCVQ global. En fait, chaque itération est un processus de codage TCVQ avec de nouveaux vecteurs-code optimisés, obtenus de l'itération précédente. En utilisant cette variante d'optimisation, l'algorithme sera dénoté par algorithme OTCVQ (Optimized Trellis Coded Vector Quantization). Des exemples de résultats de simulation obtenus lors du codage de sources sans mémoire gaussiennes, en utilisant nos encodeurs TCVQ avec dictionnaires optimisés, sont tabulés ci-dessous.

Table 2 : Performances de la TCVQ optimisée (à des débits fractionnaires) d'une source Gaussienne.

Débit (bpe)	Dim. k	Nombre d'états							VQ LBG	$D(R)$
		4	8	16	32	64	128	256		
0.66	6	3.41	3.45	3.47	3.48	3.49	3.52	3.53	3.03	4.01
0.75	4	3.81	3.85	3.87	3.89	3.93	3.96	3.97	3.35	4.51
0.80	5	4.08	4.14	4.16	4.17	4.21	4.23	4.25	3.70	4.82

En comparant ces résultats avec ceux donnés dans la table 1, nous remarquons bien l'amélioration des performances apportée par l'optimisation des dictionnaires de la TCVQ.

3. CODAGE DES LSF PAR TCVQ OPTIMISEE

En utilisant la technique OTCVQ décrite précédemment, le schéma d'encodage LSF-OTCVQ et ses performances sont présentés dans cette section. Pour des applications de codage de la parole, la TCVQ est utilisée en mode bloc, où chaque bloc correspond à un vecteur LSF de taille 10. Dans notre travail, des dictionnaires 2-D sont utilisés pour coder les vecteurs LSF. Ainsi, chaque étage dans le treillis est associé à 2 dimensions du vecteur LSF. Par conséquent, le treillis du LSF-OTCVQ est composé de 5 étages. Puisque les paramètres LSF ont des moyennes et des variances disparates, différents dictionnaires doivent être utilisés pour chaque étage. Ainsi, 5 dictionnaires 2-D étendus correspondant aux 5 étages du treillis sont nécessaires pour coder un vecteur LSF. Ces dictionnaires sont conçus d'abord par l'algorithme LBG-VQ [11]. Une fois que la structure du treillis, les sous-ensembles et l'étiquetage des branches ont été conçus, les dictionnaires sont ensuite optimisés pour la base de données (vecteurs LSF) d'apprentissage, en utilisant l'algorithme OTCVQ.

Les performances du quantificateur sont évaluées par la distorsion spectrale moyenne SD (Spectral Distorsion) qui est souvent utilisée comme mesure objective de la performance d'encodage des paramètres LSF. Calculée discrètement sur une largeur de bande limitée, l'expression de la SD pour une trame i est donnée en décibels par [5]:

$$SD_i = \sqrt{\frac{1}{n_1 - n_0} \sum_{n=n_0}^{n_1-1} \left[10 \log_{10} \frac{S(e^{j2\pi n / N})}{\hat{S}(e^{j2\pi n / N})} \right]^2} \qquad (1)$$

Pour un signal de parole échantillonné à 8 kHz avec une largeur de bande de 3 kHz, une FFT de N=256 points est utilisée pour calculer les spectres de puissance, originaux et quantifiés, $S(e^{j2\pi n/N})$ et $\hat{S}(e^{j2\pi n/N})$ du filtre de synthèse LPC de la $i^{\text{ème}}$ trame du signal de parole. Ainsi, la SD est calculée avec une résolution de 31.25 Hz par échantillon. Dans [3], Paliwal et Atal ont établi que la SD n'est pas suffisante pour mesurer seule la qualité perçue. Selon leurs résultats, un codage-LPC transparent est obtenu si les trois conditions suivantes sont maintenues: 1)- la SD moyenne est d'environ 1 dB, 2)- le pourcentage des trames "outliers frames" ayant une SD entre 2 et 4 dB est moins de 2% et 3)- aucune trame ne doit avoir une SD supérieure à 4 dB.

La base de données parole utilisée dans les simulations se compose d'environ 43 minutes de parole prise de la base de données TIMIT [12]. Les signaux de parole sont d'abord filtrés passe-bas à une fréquence 3.2 KHz, puis sous échantillonnés à 8 KHz. Pour construire la base de données des vecteurs LSF, une analyse LP, d'ordre 10 par la méthode d'autocorrélation, est effectuée sur chaque fenêtre d'analyse de 30 ms (pondérée par la fenêtre de Hamming). Une partie de la base de données (75000 vecteurs LSF) est utilisée pour l'apprentissage et la partie restante, de 11262 vecteurs LSF (différente de la base d'apprentissage), est utilisée pour les tests.

Nous évaluons, à présent, les performances de notre système OTCVQ utilisé pour la quantification des LSF. Deux mesures de distorsion différentes ont été testées séparément dans la conception et le fonctionnement du LSF-OTCVQ. Les résultats de simulation reportés ici ont été obtenus en utilisant des treillis d'Ungerboeck à 4 états. Pour différents débits de codage, les performances de l'encodeur optimisé LSF-OTCVQ, en terme de SD moyenne et de "outliers", sont montrées dans la table 3. Ces résultats ont été obtenus en utilisant une mesure de distance euclidienne non pondérée dans l'étape de conception et de fonctionnement de l'encodeur.

Table 3 : Performances de l'encodeur LSF-OTCVQ en fonction du débit de codage.

Bits/ trame	SD Moy. (dB)	Outliers (en %)	
		2-4 dB	> 4 dB
24	1.34	7.04	0.03
25	1.24	3.97	0.03
26	1.18	3.01	0.02
27	1.14	2.95	0.02
28	1.04	1.60	0.01

Ces résultats montrent que l'encodeur LSF-OTCVQ, avec la mesure de distance Euclidienne non pondérée, a besoin d'environ 28 bits/trame (ou de plus) afin de réaliser un codage de quantification transparente.

4. APPLICATION D'UNE MESURE DE DISTANCE PONDEREE DANS L'ENCODEUR OTCVQ

Afin d'obtenir une quantification transparente à des débits plus bas, une autre mesure de distorsion plus appropriée a été choisie. Il s'agit de la mesure de distance euclidienne pondérée. Basé sur les principales propriétés des paramètres LSF, certaines formules de distances pondérées ont été proposées pour le codage des LSF [3, 13]. Si f et $\hat{f}$ sont respectivement les deux vecteurs original et quantifié des paramètres LSF, alors la distance euclidienne carrée pondérée entre ces deux vecteurs est donnée par [3, 13] :

$$d(f, \hat{f}) = \sum_{i=1}^{10} c_i w_i (f_i - \hat{f}_i)^2 \qquad (2)$$

où c_i et w_i représentent respectivement les poids fixe et variable assignés au $i^{\text{ème}}$ coefficient LSF. Plusieurs fonctions de pondération ont été définies pour calculer le vecteur des poids variables $w = [w_1, \ldots, w_{10}]$. Dans notre conception, nous avons utilisé la fonction de pondération proposée dans [14]. Elle définie par :

$$w_i = \frac{1}{f_i - f_{i-1}} + \frac{1}{f_{i+1} - f_i}, \quad (3)$$

avec $f_0 = 0$ et $f_{11} = 0.5$.

Le vecteur additionnel de poids constants $c = [c_1,\ldots, c_{10}]$ a été introduit dans la formule afin que les LSF des basses fréquences soit pondérés plus que les autres. Ce vecteur est déterminé expérimentalement [3] :

$$ci = \begin{cases} 1.0, & for \ \ 1 \le i \le 8 \\ 0.8, & for \ \ i = 9 \\ 0.4, & for \ \ i = 10 \end{cases} . \quad (4)$$

Pour différents débits de codage, les performances du système d'encodage LSF-OTCVQ, utilisant une mesure de distance euclidienne pondérée sont tabulées ci-dessous.

Table 4 : Performances de l'encodeur LSF-OTCVQ en utilisant une mesure de distance pondérée

Bits/ Trame	SD Moy. (dB)	Outliers (en %)	
		2-4 dB	> 4 dB
24	1.29	5.26	0.02
25	1.19	2.99	0.00
26	1.15	2.72	0.00
27	1.07	1.90	0.00
28	0.98	1.10	0.00

En comparant ces résultats à ceux donnés dans la table 3, nous pouvons constater que la mesure de distance pondérée améliore les performances de l'encodeur LSF-OTCVQ en termes de SD moyenne et de nombre de trames "outliers". Nous avons besoin ici de 27 bits/trame pour obtenir une qualité de quantification transparente.

5. CONCLUSION

Dans ce travail, un système basé sur la quantification vectorielle codée par treillis a été appliqué avec succès dans le codage efficace des paramètres LSF. En utilisant une mesure de distance pondérée dans la conception et l'opération de notre encodeur LSF-OTCVQ, une quantification de qualité transparente peut être réalisée à des débits plus bas. Comparé à l'encodeur conçu avec une distance non pondérée, l'LSF-OTCVQ avec la distance pondérée peut diminuer le débit d'environ 1-2 bits/trame, tout en maintenant des performances comparables.

BIBLIOGRAPHIE

[1] Kleijn (W.B.), Paliwal (K.K.), Speech coding and synthesis, *Elsevier Science B.V.*, 1995.

[2] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", Journ. of Acoust. Society America, vol.57, p.535, 1975.

[3] K. K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", IEEE Transactions on Speech and Audio Processing, vol. 1, no. 1, pp. 3-14, Jan. 1993.

[4] F. K. Soong and B. H. Juang, "Optimal quantization of LSP parameters," Proc. IEEE Int. Conf. Acous., Speech Signal Processing, New York, pp. 394-397, April 1988.

[5] W. F. LeBlanc, B. Bhattacharya, S. A. Mahmoud & V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding", IEEE Trans. Speech and Audio Processing, Vol. 1, No. 4, pp. 373-385, October 1993.

[6] M.W. Marcellin & T.R. Fischer, "Trellis coded quantization of memoryless and Gauss-markov sources ", IEEE Trans. on Communication, Vol. 38, pp. 83-93, January 1990.

[7] T.R. Fischer, M.W. Marcellin & M. Wang, "Trellis-coded vector quantization", IEEE Trans. on Information Theory, Vol.37, pp.1551-1566, November 1991

[8] H.S. Wang & N. Moayeri, "Trellis coded vector quantization", IEEE Trans. on Communication, Vol. 40, pp. 1273-1276, August 1992.

[9] G. Ungerboeck, "Trellis-coded modulation with redundant signal sets, Part I and II", IEEE Commun. Magazine, Vol.25, pp.5-21, Feb. 1987.

[10] M. BOUZID, "Trellis Coded Vector quantization", Technical research report, speech coding team, Centre de Développement des Technologies avancées (CDTA), Alger, 2001.

[11] Y.Linde, A.Buzo & R.M.Gray, "An Algorithm for Vector Quantization Design", IEEE Trans. on Comm., Vol. COM-28, pp.84-95, Jan. 1980.

[12] J. S. Garofolo et al., "DARPA TIMIT Acoustic-phonetic Continuous Speech Database", NIST, Gaithersburg, October 1988.

[13] F. Lahouti, & A.K. Khandani, "Quantization of LSF Parameters Using A Trellis Modeling", IEEE Trans. on Speech and Audio Processing, Vol. 11, Issue 5, pp. 400-412, Sep. 2003.

[14] R. Laroia, N. Phamdo & N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers", Proc. IEEE Int. Conference Acoustic Speech and Signal Processing, pp. 641-644, May 1991.

Evaluation d'un système de synthèse 3D de Langue française Parlée Complétée

G. Gibert, G. Bailly, F. Elisei

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex, France
{gibert, bailly, elisei}@icp.inpg.fr

ABSTRACT

This paper presents the virtual speech cuer built in the context of the ARTUS project aiming at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then incrusted - on demand and at the reception - in the original broadcast as an alternative to subtitling. The paper presents the multimodal text-to-speech synthesis system and the first evaluation performed by deaf users.

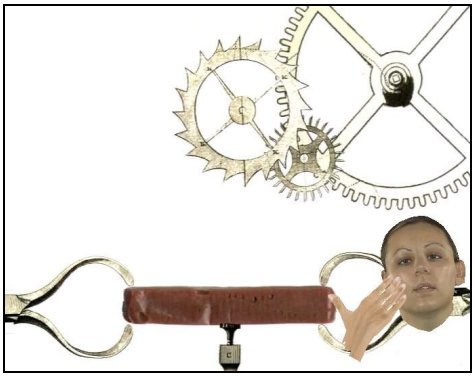


Figure 1: Incrustation du clone ARTUS dans un documentaire produit par la chaîne ARTE.

1. INTRODUCTION

Les personnes sourdes ou malentendantes dépendent grandement de la lecture labiale qui est basée sur l'information visuelle délivrée par les lèvres et le visage. Cependant, la lecture labiale seule est insuffisante dû à un manque d'information sur le point de l'articulation de la langue, des modes d'articulation (nasalité, voisement) et à la similarité de certaines formes labiales pour certains phonèmes (aussi appelés sosies labiaux tels que [u] *vs* [y]). Dans tous les cas, même le meilleur décodeur ne peut identifier plus de 50% de phonèmes dans des syllabes sans sens [16] ou dans des mots ou des phrases [5]. Le système de codage de la Langue française Parlée Complétée a été construit pour compléter la lecture labiale. Développé par Cornett [7, 9] et adapté depuis à plus de 50 langues [8], ce système est basé sur l'association de l'articulation faciale avec des clés formées par la main. Une clé est caractérisée par une position sur le visage (déterminant un sous-ensemble de voyelles) et une forme de main (déterminant un sous-ensemble de consonnes). Le code LPC pour les consonnes est représenté sur la Figure 2. De nombreuses études ont montré le gain d'intelligibilité apporté par ce codage comparé à la lecture labiale seule [15, 19] et son efficacité dans l'apprentissage de la langue écrite et orale [13, 14].

Des travaux sont consacrés à l'étude de la perception du code LPC et à sa production [1] mais peu de travaux s'attachent à la synthèse de celui-ci [voir les systèmes basés sur des règles dans 2, 10]. Nous allons décrire le système de synthèse multimodal produisant du code LPC à partir du texte et la première campagne d'évaluation auprès de personnes sourdes et malentendantes.

2. LE SYSTEME DE SYNTHESE MULTIMODAL

Le système de synthèse 3D de code LPC développé dans le cadre du projet ARTUS convertit une série de sous-titrage télétexte en un flux de paramètres d'animation pour les mouvements de la tête, du visage, du bras et de la main et produit également un signal acoustique. Les modèles de contrôle, de forme et d'apparence ont été déterminés à l'aide de plusieurs enregistrements multimodaux d'une locutrice oralisant et codant LPC.

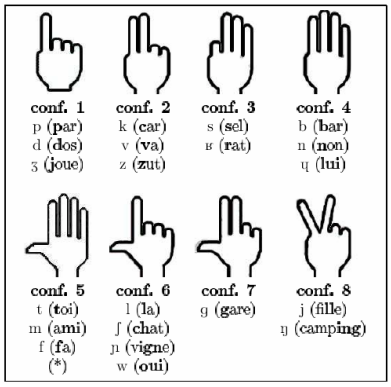


Figure 2 : Système de codage LPC pour les consonnes.

2.1. Expérimentation et modélisation

Les différentes configurations expérimentales pour enregistrer notre codeuse et capturer ses mouvements sont décrites dans [12]. Les configurations incluent (a) un système de capture de mouvements avec une bonne résolution temporelle (120Hz) et une bonne précision spatiale (0.1mm) de quelques dizaines de marqueurs rétro-réfléchissants collés sur le visage et la main de la codeuse (voir Figure 3), lors de la production de code LPC de 238 phrases; (b) un système de capture vidéo de plusieurs centaines de billes colorées collées sur le visage de notre locutrice (voir Figure 4), lors du codage de simples syllabes et

(c) des textures cylindriques de la tête, des moulages de sa main et de ses dents. A l'aide de toutes ces données, des modèles précis de forme et d'apparence de la tête et du visage ainsi que de la main de la codeuse ont été développés [11]. Ces deux types de modèles (forme et apparence) du visage et de la main sont pilotés par des paramètres quasi-articulatoires issus d'une analyse en composantes principales guidée par des connaissances articulatoires a priori.

2.2 Système de synthèse 3D de code LPC

Le système de synthèse de parole à partir du texte COMPOST [3] a été paramétré et des modules ont été ajoutés afin de générer la Langue française Parlée Complétée. Les ajouts sont les suivants:

Traitements linguistiques. Les sous-titrages ayant une ponctuation lâche voire absente, un module spécifique considère le début de chaque morceau de texte comme le début potentiel d'une phrase et abandonne les hypothèses peu probables. Pour respecter la synchronisation avec le contenu visuel, issue du télétexte originel, des marqueurs temporels sont insérés au début des phrases détectées. Ensuite, le module rythmique adapte la durée des pauses inter-phrases pour attendre ces rendez-vous.

Prosodie. Même si les codeurs LPC sont capables de minimiser l'impact de l'ajout d'un geste modal sur le débit de parole, le codage d'une consonne isolée (dans un cluster de consonnes ou dans une coda) impose un débit de parole plus faible et une hyperarticulation des syllabes complexes. L'intonation est également affectée. Le modèle prosodique SFC [4] a donc par conséquent été entraîné sur les données expérimentales. Avec le modèle ainsi appris, trois émissions télévisuelles ont été interprétées : seules quatre phrases n'ont pas été prononcées dans le temps imparti (retard moyen de 120 ms). Pour note, les règles gouvernant la création des sous-titres considèrent uniquement le nombre de lettres d'un groupe et non le temps de lecture.

Synthèse par concaténation d'unités multimodales. Le son et les mouvements de synthèse sont produits par sélection, lissage et concaténation d'unités multimodales multi-représentées. Deux types de segments sont considérés et synchronisés suivant des repères temporels acoustiques et gestuels déduits de règles spécifiques [12]: les "polysons" capturant le signal acoustique et les mouvements faciaux entre deux cibles acoustiques stables (les sons tels que les glides sont inclus dans des unités plus grandes) et les "diclés" qui contiennent les mouvements du bras, de la main et de la tête entre deux clés successives. Les mouvements de la tête contribuent, dans le cas de notre codeuse, significativement à la constriction main/visage : même si la main contribue pour la plus grande partie du mouvement menant la main à une position par rapport au visage, la tête effectue en moyenne à 16,43% de la distance à parcourir. Notons qu'une telle contribution des gestes posturaux à la structuration du discours a déjà été rapportée pour les signeurs natifs [6].

Les segments multi-représentés sont sélectionnés par un algorithme classique de programmation dynamique qui utilise un coût de sélection et un coût de concaténation. Le coût de concaténation prend en compte la contribution relative de chaque paramètre d'animation par rapport à la variance du mouvement total expliqué.

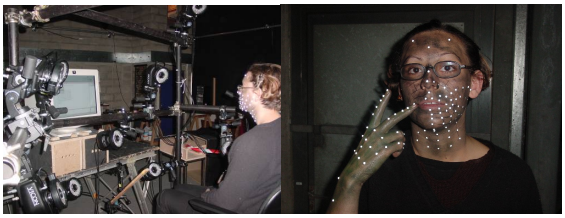


Figure 3 : Capture de mouvement avec un système Vicon® (12 caméras, 120Hz, 50 marqueurs sur la main et 63 sur le visage)

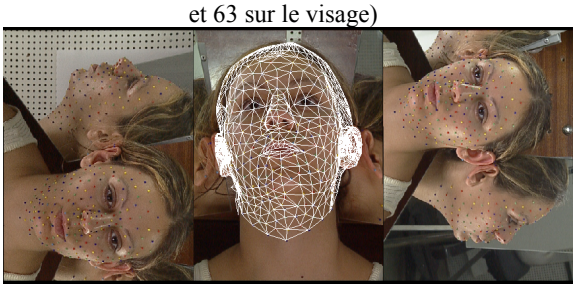


Figure 4 : Capture vidéo avec 247 billes collées sur le visage de la codeuse.

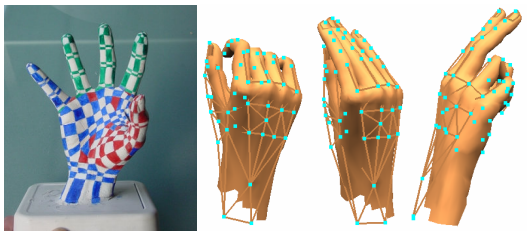


Figure 5 : A gauche : un moulage de la main. A droite : le modèle de main déduit est contrôlé par les données issues de la capture de mouvements.

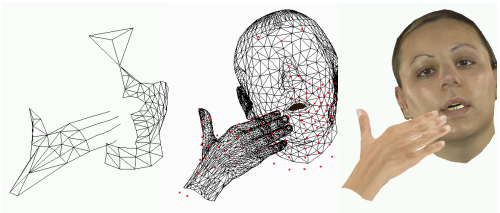


Figure 6 : Le clone du projet ARTUS. De la capture du mouvement à une animation vidéo-réaliste.

2.3 Animation vidéoréaliste

Un modèle d'apparence vidéo-réaliste a été développé pour la main de notre codeuse en utilisant des moulages de la main et la technique de *skinning* [20] de l'infographie (voir Figure 5). Des modèles génériques de lèvres, du crâne, de dents et d'yeux ont été adaptés à la morphologie du sujet. Le résultat de

ces procédures est la création d'un modèle de forme haute définition texturé par un mélange de textures cylindriques et contrôlé par un ensemble de paramètres quasi-articulatoires. Le clone vidéoréaliste résultant est représenté sur la Figure 6.

3. EVALUATION

Une première série d'expériences a été conduite afin d'évaluer l'intelligibilité de notre codeur face à des personnes malentendantes ou sourdes utilisant le code LPC. La première campagne d'évaluation est dédiée à l'intelligibilité segmentale tandis que la seconde est consacrée à la compréhension.

3.1 Intelligibilité segmentale

Paires minimales. Le test développé pour l'occasion est une adaptation du test de rhyme adapté pour le français par Peckels et Rossi [17]: les paires minimales ne testent pas, dans notre cas, les traits acoustiques mais les traits gestuels. Une liste de mots appariés de type CVC a été construite pour tester systématiquement les paires de consonnes en position initiale qui ne différent que dans la forme de main associée. Nous avons choisi toutes les paires dans chacun des 8 sous-ensembles codant les consonnes en LPC et qui étaient visuellement très proches [18]. Les voyelles centrales ont été choisies de telle sorte que toutes les positions par rapport au visage soient présentes et les consonnes finales ont été choisies afin de tester la capacité du système à gérer correctement la coarticulation. Comme toutes les paires minimales n'ont pu être générées dans tous les contextes vocaliques, nous obtenons une liste finale de 196 mots.

Conditions. Les stimuli par paires minimales sont présentés aléatoirement dans les deux ordres. La modalité lecture labiale seule est testée en premier. La modalité incluant le code LPC est présentée dans un deuxième temps afin de réserver les ressources cognitives pour la tâche la plus difficile i.e. la première tâche.

Stimuli. Pour la modalité de présentation « lecture labiale », afin d'éviter la présentation d'une tête complètement statique qui pourrait sembler non naturelle, nous avons divisé par 10 les mouvements de tête fournis par le système de synthèse. Aucune modification des mouvements segmentaux ou suprasegmentaux n'a été effectuée de manière à hyper-articuler.

Sujets. Les sujets sont au nombre de huit, ils sont sourds ou malentendants ayant appris la Langue française Parlée Complétée dès l'âge de 3 ans.

Résultats. Le taux de reconnaissance moyen pour la modalité « lecture labiale » est de 52.36%. Ce taux signifie que les paires proposées ne sont pas distinguables. Il revient donc au même de répondre au hasard. Il s'agit d'un premier résultat qui confirme que nos formes labiales entre sosies labiaux sont assez proches pour être confondues.

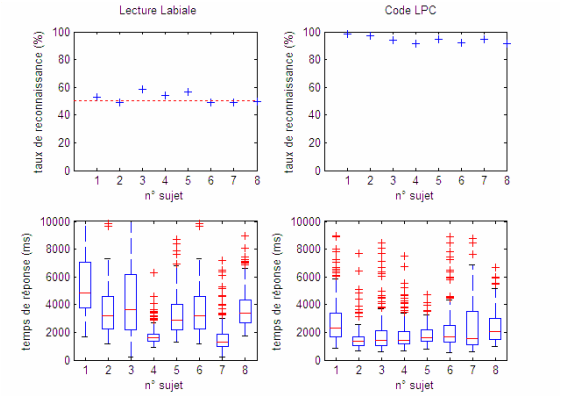


Figure 7 : Taux de reconnaissance et temps de réponse des 8 sujets pour les deux modalités (lecture labiale et lecture labiale + code LPC).

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	l	j
d	48	18	-	18	20	-	-	-	-	-	-	-	-	-	-	-	-
t	29	71	-	-	23	-	-	-	-	-	16	-	-	-	-	19	2
n	-	-	69	10	17	-	-	-	-	-	7	-	-	-	-	11	6
z	11	-	4	51	14	-	-	-	-	-	-	-	-	-	-	-	-
s	22	15	16	22	77	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	15	25	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	15	41	24	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	24	16	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	57	17	9	-	-	-	-	-	5
ʃ	-	-	-	-	-	-	-	-	17	61	12	14	-	-	-	-	-
g	-	14	7	-	-	-	-	-	15	19	115	9	-	-	17	-	4
k	-	-	-	-	-	-	-	-	-	13	11	45	-	-	-	-	3
v	-	-	-	-	-	-	-	-	-	-	-	-	17	15	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	18	14	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	17	-	-	-	51	17	3
l	-	22	27	-	-	-	-	-	-	-	-	-	-	-	24	47	-
j	-	4	3	-	-	-	-	-	13	-	1	2	-	-	7	-	26

Figure 8 : Matrice de confusion de la consonne initiale pour la modalité « lecture labiale ».

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	l	j
d	102	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
t	2	155	-	-	-	-	-	-	-	-	1	-	-	-	-	2	-
n	-	-	99	-	17	-	-	-	-	-	1	-	-	-	-	3	-
z	2	-	-	74	4	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	1	-	151	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	40	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	1	74	5	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	5	35	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	85	2	1	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	-	3	100	1	-	-	-	-	-	-
g	-	3	1	-	-	-	-	-	3	6	180	2	-	-	5	-	-
k	-	-	-	-	-	-	-	-	-	4	6	62	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	-	32	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	3	29	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-
l	-	1	1	-	-	-	-	-	-	-	-	-	-	-	1	117	-
j	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55

Figure 9 : Matrice de confusion de la consonne initiale pour la modalité « lecture labiale + code LPC ».

Le taux de reconnaissance moyen pour la modalité « lecture labiale + code LPC » est de 94.26%. La différence de taux de reconnaissance entre les deux modalités montre que notre codeur LPC apporte une information significative en terme de mouvements de main. On peut voir sur les matrices de confusion (Figure 8 et Figure 9), les erreurs faites par les sujets. Pour le groupe des consonnes bilabiales (encadré rouge), la consonne /p/ n'est pas reconnue dans la modalité « lecture labiale » (25 fois sur 40 elle est reconnue comme un /b/) alors qu'elle est toujours reconnue dans la modalité « lecture labiale + code LPC ».

Les temps de réponse qui sont un indice de la charge cognitive imposée aux sujets pour effectuer la tâche de discrimination entre les paires sont significativement différents (ANOVA à un facteur à mesures répétées $F(1,3134)=7.5$, $p<0.01$). Il est ainsi plus aisé pour les sujets de répondre à la tâche incluant le code LPC qu'à la tâche « lecture labiale ». Le gain est donc double, en termes de reconnaissance et en termes de charge cognitive nécessaire.

3.2 Compréhension

Afin d'évaluer la compréhension globale de notre système, nous avons demandé au sujets de l'étude précédente de visualiser un reportage de l'émission *Karambolage* de la chaîne ARTE dans lequel le clone LPC est incrusté (voir Figure 1). A la fin de la séance, nous avons demandé aux sujets de répondre à un questionnaire. Ce questionnaire se compose de 10 questions portant tant sur les informations apportées par la vidéo originale que par le clone ARTUS.
Le nombre moyen de réponses correctes est de 3 sur 10. Les sujets rapportent comprendre des mots isolés mais pas l'ensemble du discours. Cette observation est également constatée si l'on présente la vidéo de la codeuse incrustée. Une explication de ces résultats se trouve dans la complexité de la tâche proposée (rythme élevé, manque de marqueurs « prosodiques », etc.).

CONCLUSIONS ET PERSPECTIVES

L'observation et l'enregistrement d'une codeuse en action nous a permis de développer un système complet de synthèse LPC 3D à partir du texte. Les résultats préliminaires des tests perceptifs appliqués à ce système soulignent l'énorme gain en intelligibilité apporté par notre système. Cette série de tests doit se poursuivre sur un plus grand nombre de sujets pour pouvoir généraliser les résultats et quantifier plus finement la charge cognitive imposée aux sujets. Elle est actuellement complétée par des tests qui comparent l'usage du clone par rapport au télétexte à l'aide d'un dispositif oculométrique.

REMERCIEMENTS

Nous tenions à remercier Yasmine Badsi, notre codeuse LPC. Nous remercions également Martine Marthouret, Marie-Agnès Cathiard, Denis Beautemps et Virginie Attina pour leur aide dans l'élaboration des tests perceptifs. Nous n'oublions pas les sujets qui ont bien voulu participer.

BIBLIOGRAPHIE

[1] Attina, V. (2006) *La Langue française Parlée Complétée (LPC) : Production et Perception*. PhD Thesis. Institut National Polytechnique: Grenoble - France.

[2] Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (2004) *A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer.* Speech Communication, **44**: p.197-214.

[3] Bailly, G. and Alissali, M. (1992) *COMPOST: a server for multilingual text-to-speech system.* Traitement du Signal, **9**(4): p.359-366.

[4] Bailly, G. and Holm, B. (2005) *SFC: a trainable prosodic model.* Speech Communication, **46**(3-4): p.348-364.

[5] Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (2000) *Speech perception without hearing.* Perception & Psychophysics, **62**: p.233-252.

[6] Brentari, D. (1999) *A prosodic model of sign language phonology*. Boston, MA: MIT Press.

[7] Cornett, R.O. (1967) *Cued Speech.* American Annals of the Deaf, **112**: p.3-13.

[8] Cornett, R.O. (1988) *Cued Speech, manual complement to lipreading, for visual reception of spoken language.* Principles, practice and prospects for automation. Acta Oto-Rhino-Laryngologica Belgica, **42**(3): p.375-384.

[9] Cornett, R.O. (1982) *Le Cued Speech*, in *Aides manuelles à la lecture labiale et perspectives d'aides automatiques*, F. Destombes, Editor. Centre scientifique IBM-France: Paris.

[10] Duchnowski, P., Lum, D.S., Krause, J.C., Sexton, M.G., Bratakos, M.S., and Braida, L.D. (2000) *Development of speechreading supplements based on automatic speech recognition.* IEEE Transactions on Biomedical Engineering, **47**(4): p.487-496.

[11] Elisei, F., Bailly, G., Gibert, G., and Brun, R. (2005) *Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis*. in *Auditory-Visual Speech Processing Workshop*. Vancouver, Canada

[12] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., and Brun, R. (2005) *Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech.* Journal of Acoustical Society of America, **118**(2): p.1144-1153.

[13] Leybaert, J. (2000) *Phonology acquired through the eyes and spelling in deaf children.* Journal of Experimental Child Psychology, **75**: p.291-318.

[14] Leybaert, J. (2003) *The role of Cued Speech in language processing by deaf children: an overview*. in *Auditory-Visual Speech Processing*. St Jorioz - France. p.179-186.

[15] Nicholls, G. and Ling, D. (1982) *Cued Speech and the reception of spoken language.* Journal of Speech and Hearing Research, **25**: p.262-269.

[16] Owens, E. and Blazek, B. (1985) *Visemes observed by hearing-impaired and normal-hearing adult viewers.* Journal of Speech and Hearing Research, **28**: p.381-393.

[17] Peckels, J.P. and Rossi, M. (1973) *Le test de diagnostic par paires minimales. Adaptation au francais du 'Diagnostic Rhyme Test' de W.D. Voiers.* Revue d'Acoustique, **27**: p.245-262.

[18] Summerfield, Q. (1991) *Visual perception of phonetic gestures*, in *Modularity and the motor theory of speech perception*, I.G. Mattingly and M. Studdert-Kennedy, Editors. Lawrence Erlbaum Associates: Hillsdale, NJ. p. 117-138.

[19] Uchanski, R., Delhorne, L., Dix, A., Braida, L., Reed, C., and Durlach, N. (1994) *Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech.* Journal of Rehabilitation Research and Development, **31**: p.20-41.

[20] Woodward, C.D. (1988) *Skinning techniques for interactives B-spline surface interpolation.* Computer-Aided Design, **20**(8): p.441- 451.

Modélisation B-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère MDL

Damien Lolive, Nelly Barbot, Olivier Boëffard

IRISA / Université de Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
damien.lolive,nelly.barbot,olivier.boeffard@irisa.fr
http ://www.irisa.fr/cordial/

ABSTRACT

This article describes a new approach to estimate F_0 curves using a B-Spline model characterized by a knot sequence and associated control points. The free parameters of the model are the number of knots and their location. The free-knot placement, which is a NP-hard problem, is done using a global MLE within a simulated-annealing strategy. The optimal knots number estimation is provided by MDL methodology. Two criteria are proposed considering control points as real coefficients with variable precision. They differ on the precision used. Experiments are conducted in a speech processing context on a 7000 syllables french corpus. We show that a variable precision criterion gives good results in terms of RMS error (0.42Hz) as well as in terms of B-spline freedom number reduction (63% of the full model).

1. INTRODUCTION

Les technologies liées au traitement de la parole font largement appel aux modèles d'intonation. Notamment, la synthèse de la parole à partir du texte, TTS - Text-to-Speech systems -, ne peut se dispenser de tels modèles de manière à prédire des contours mélodiques à partir du texte et du style oratoire. Plus récemment, ces modèles ont été utilisés afin de déterminer une séquence optimale d'unités acoustiques prenant en compte des caractéristiques prosodiques[7]. Bien que l'intonation résulte d'une combinaison de nombreux facteurs linguistiques, cet article traite exclusivement du paramètre acoustique reconnu pour être le principal facteur de perception de la prosodie, à savoir la fréquence fondamentale ou F_0. Les contours de F_0, extraits du signal de parole, résultent de l'évolution au cours du temps de la vibration des cordes vocales. Une littérature importante traite de la modélisation de ces contours. On peut citer en particulier les modèles MoMel [4], Tilt [9], ainsi que les travaux de Sakai et Glass [8] qui utilisent des fonctions splines régulières.

Dans cet article, nous proposons une stylisation B-spline qui intègre une notion de régularité locale sans affecter l'approximation globale de la courbe F_0, estimée selon un critère des moindres carrés. Pour conduire des expériences d'évaluation, nous avons dû faire le choix d'une unité prosodique minimale. Nous choisi la syllabe comme support minimal d'un contour mélodique. Bien entendu, la méthodologie de stylisation proposée ne fait aucune hypothèse de cette nature. Dans [1], une comparaison entre les capacités de modélisation de F_0 par des splines et des B-splines est présentée selon un critère des moindres carrés. Les paramètres du modèle B-spline sont le nombre de nœuds et leur placement. Un placement libre des nœuds permet de suivre les irrégularités des contours. Il est mis en œuvre par un algorithme de type Monte-Carlo (recuit simulé) afin de contourner la difficulté combinatoire de ce problème. Notre principal objectif est alors d'estimer le nombre optimal de nœuds afin de déterminer une classe de courbes B-splines parcimonieuse. Le cadre que nous avons choisi est celui du critère de longueur de description minimale, Minimum Description Length - MDL -, qui offre un compromis efficace entre la précision du modèle et le nombre de ses paramètres.

Dans la section 2, on introduit le modèle B-spline et ses paramètres estimés dans la section 3 au sens des moindres carrés. Dans la section 4, on présente les critères MDL pour optimiser le nombre de paramètres du modèle. Dans la section 5, le protocole expérimental est décrit et les résultats sont donnés dans la section 6.

2. MODÉLISATION B-SPLINE

Dans ce paragraphe, on présente les fonctions splines et leur généralisation aux courbes B-splines qui ont la capacité de modéliser des courbes ouvertes ayant des régularités locales variables.

Soit $[a, b]$ un intervalle que l'on subdivise en $l + 1$ sous-intervalles : $a = t_m < \ldots < t_{m+l+1} = b$. Une spline de degré m associée à (t_i) est une fonction polynomiale de degré m sur chaque sous-intervalle et de classe $\mathcal{C}^{m-1}$ sur $[a, b[$. Les l points de jonction sont appelés des nœuds internes. L'ensemble de ces splines forme un espace vectoriel de dimension $m + l + 1$.

Dans [4], l'algorithme MoMel a pour but de fournir une stylisation de contours mélodiques à l'aide d'une spline quadratique. Pour cela, il estime n points cibles constituant les points stationnaires d'une spline de degré 2. Les points t_i correspondent aux abscisses des n points cibles et à leurs $(n - 1)$ points médians. Ainsi, la spline quadratique est associée à $l = 2n - 3$ nœuds internes et est entièrement déterminée par les $2n$ contraintes dûes aux n points à interpoler avec une tangente horizontale.

L'espace vectoriel des splines de degré m associées à (t_i) admet une base de fonctions B-splines. Pour les définir, on pose $t_0 = \ldots = t_m = a$ et $t_{m+l+1} = t_{2m+l+1} = b$, appelés nœuds externes, et on note $\mathbf{t}$ le vecteur contenant les nœuds internes et externes. Les B-splines sont alors définies par récurrence : au degré 0, pour i de 0 à $2m + l$

$B_0^i = \mathbf{1}_{[t_i, t_{i+1}[}$ et au degré m, pour i de 0 à $m+l$

$$B_m^i(t) \;=\; \frac{t-t_i}{t_{i+m}-t_i} B_{m-1}^i(t) + \frac{t_{i+m+1}-t}{t_{i+m+1}-t_{i+1}} B_{m-1}^{i+1}(t)$$

où les quotients sont nuls si $t_i = t_{i+m}$ ou $t_{i+1} = t_{i+m+1}$. Ajoutons que les B-splines sont positives et vérifient

$$\forall t \in [a,b[\,, \quad \sum_{i=0}^{m+l} B_m^i(t) \;=\; 1\,. \tag{1}$$

Par conséquent, une fonction spline de degré m s'écrit comme une combinaison linéaire de fonctions B-splines de même degré. La généralisation des splines aux courbes B-splines autorise des nœuds internes non distincts et leur multiplicité désigne le nombre de fois où ils apparaissent dans $\mathbf{t}$. Ainsi, une courbe B-spline de degré m associée à un vecteur nœud $\mathbf{t}$ s'écrit comme une combinaison linéaire des fonctions B-splines de degré m dont les coefficients $c_0, \ldots, c_{m+l}$ sont appelés points de contrôle.

On peut représenter matriciellement les points d'une courbe B-spline. Soit $x_0, \ldots, x_{N-1}$ une suite de valeurs de $[a,b]$, on définit la matrice $\mathbf{B}$ qui a $B_m^i(x_j)$ comme (j,i)-ème élément, et on a, pour tout j de 0 à $N-1$

$$g_{\theta^l}(x_j) \;=\; \sum_{i=0}^{m+l} c_i B_m^i(x_j) \;=\; (\mathbf{Bc})_j \tag{2}$$

où $\mathbf{c}$ désigne le vecteur colonne contenant les points de contrôle et $\theta^l = (t_{m+1}, \ldots, t_{m+l}, c_0, \ldots, c_{l+m})$.

Le principal effet des nœuds sur la courbe g_{θ^l} est lié à leur multiplicité. Soit t_i un nœud interne, plus sa multiplicité m_i est élevée et moins la courbe B-spline est régulière au point t_i. Plus précisément, si g_{θ^l} est de classe $\mathcal{C}^{m-1}$ entre deux nœuds consécutifs, elle est de classe $\mathcal{C}^{m-m_i}$ au nœud t_i. Quant aux points de contrôle, ils ont un impact local. En effet, si l'on fait varier c_i, cela modifie le terme $c_i B_m^i$ dans (2) et B_m^i étant nulle en dehors de $[t_i, t_{i+m+1}[$, seul le segment de courbe correspondant est affecté.

3. ESTIMATION DES PARAMÈTRES

Soit $\{(x_j, y_j), j = 0, \ldots, N-1\}$ un ensemble de mesures d'un contour mélodique, où (x_j) forme une suite croissante. On cherche à déterminer la courbe B-spline g_{θ^l} de degré m minimisant l'erreur quadratique moyenne avec les observations précédentes. Dans cette partie, on suppose le nombre l de nœuds internes connu, et on pose $t_0 = t_m = x_0$ et $t_{m+l+1} = t_{2m+l+1} = x_{N-1}$. On optimise dans un premier temps les points de contrôle, puis on estime la localisation des nœuds internes.

3.1. Les points de contrôle

Dans ce paragraphe, le vecteur nœud $\mathbf{t}$ est supposé connu et on détermine la courbe B-spline g_{θ^l} de degré m, i.e. ses points de contrôle, minimisant l'erreur quadratique avec les données. On note $\mathbf{y}$ le vecteur colonne de coordonnées y_j. D'après (2), on estime les points de contrôle tels que

$$\widehat{\mathbf{c}} \;=\; \arg\min_{\mathbf{c}} ||\mathbf{y} - \mathbf{Bc}||_2^2 \;=\; \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{y} \tag{3}$$

si $\mathbf{B}^T\mathbf{B}$ est inversible, i.e. lorsque les colonnes de $\mathbf{B}$ sont linéairement indépendantes. Pour cela, aucune multiplicité de nœud ne doit être supérieure à $m+1$, sous peine d'avoir une colonne nulle. De plus, le nombre N de lignes de cette matrice doit être nettement supérieur au nombre de colonnes afin de bien différencier chaque colonne de $\mathbf{B}$, i.e.

$$N \;\gg\; m+l+1\,. \tag{4}$$

3.2. Placement des nœuds

On souhaite à présent déterminer un placement optimal des nœuds internes. Pour cela, on introduit le critère du maximum de vraisemblance et on choisit une stratégie de type Monte-Carlo (recuit simulé) pour déterminer une solution $\widehat{\mathbf{t}}$ à ce problème NP-difficile. On considère que les nœuds se situent à des endroits d'observations, ils peuvent alors être représentés par un entier entre 0 et N-1.

On note e_j l'erreur commise entre l'observation y_j à l'instant x_j et sa modélisation $g_{\theta^l}(x_j)$. Afin de simplifier les calculs, on suppose $e_0, \ldots, e_{N-1}$ indépendantes et distribuées selon une loi gaussienne centrée de variance σ^2. Par conséquent, la log-vraisemblance du modèle s'écrit

$$\log p\left(\mathbf{y}; \theta^l\right) \;=\; -\frac{1}{2\sigma^2}\,||\mathbf{y} - \mathbf{Bc}||_2^2 - \frac{N}{2}\log\left(2\pi\sigma^2\right)\,. \tag{5}$$

On remarque que $\widehat{\mathbf{c}}$, défini par (3), est l'estimateur du maximum de vraisemblance pour $\mathbf{c}$. L'estimateur $\widehat{\mathbf{t}}$ du maximum de vraisemblance pour $\mathbf{t}$ vérifie

$$\widehat{\mathbf{t}} \;=\; \arg\min_{\mathbf{t}} \left|\left|\mathbf{y} - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y}\right|\right|_2^2$$

où la matrice $\mathbf{B}$ dépend du paramètre $\mathbf{t}$.

Afin de déterminer $\widehat{\mathbf{t}}$, nous avons mis en œuvre un algorithme d'optimisation globale selon une stratégie de type recuit-simulé (Simulated Annealing SA), présenté dans [1]. La relation proposée entre les paramètres du modèle et les distributions de l'algorithme SA est la suivante :

1. SA échantillonne un vecteur $\mathbf{v}$ de $\{x_1, \ldots, x_{N-2}\}^l$.
2. Un vecteur $\mathbf{t}^*$ de nœuds internes est défini en triant les coordonnées de $\mathbf{v}$.
3. Un vecteur nœud $\mathbf{t}$ est défini en ajoutant $(m+1)$ fois x_0 et x_{N-1} aux extrémités de $\mathbf{t}^*$.
4. Si deux nœuds consécutifs se trouvent dans un intervalle inférieur à 5% de $[x_0, x_{N-1}]$, ils sont fusionnés et la multiplicité du premier est incrémentée de 1.

La prochaine étape concerne l'optimisation du nombre l de nœuds internes. Avant de l'aborder, résumons les choix effectués quant au modèle B-spline. Pour l fixé, on commence par estimer un vecteur-nœud $\widehat{\mathbf{t}}$ optimal à l'aide d'un algorithme SA . À partir de la matrice $\mathbf{B}$ associée et de (3), on calcule les points de contrôle $\widehat{\mathbf{c}}$ optimaux. Par la suite, on désigne par $\widehat{\theta^l}$ le modèle estimé.

4. B-SPLINES ET MDL

Le critère de longueur de description minimale (MDL) permet d'établir un compromis entre la qualité d'un modèle $\widehat{\theta^l}$ et sa complexité l. Plus précisément, son application consiste à déterminer $\widehat{l}$ minimisant la longueur de description des données $\mathbf{y}$. D'après [3],

$$\widehat{l} \;=\; \arg\min_l L(\mathbf{y}) \;=\; \arg\min_l L\left(\widehat{\theta^l}\right) - \log_2 p\left(\mathbf{y}; \widehat{\theta^l}\right)$$

où $L\left(\widehat{\theta^l}\right)$ et $-\log_2 p\left(\mathbf{y}; \widehat{\theta^l}\right)$ désignent respectivement les longueurs de description du modèle estimé et des ob-

servations sachant ce dernier. L'expression (5) fait intervenir la variance σ^2 de l'erreur. Son estimateur selon le maximum de vraisemblance s'écrit :

$$\widehat{\sigma^2} \;=\; \arg\max_{\sigma^2} \log p\left(\mathbf{y}; \widehat{\theta^l}\right) \;=\; \frac{1}{N}\,||\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}||_2^2 \;.$$

Par conséquent, l'écart-type σ est estimé par la racine carrée de la moyenne des carrés des erreurs entre les observations et leur modélisation B-spline, également appelée erreur RMS (Root Mean Square). En injectant cette estimation dans (5), l'expression ci-dessus devient

$$L(\mathbf{y}) \;=\; L\left(\widehat{\theta^l}\right) + \frac{N}{2} + \frac{N}{2}\log_2(2\pi) + N\log_2\left(RMS\right) \;.$$

4.1. Principe de solution

Considérons la longueur de description du vecteur $\widehat{\theta^l}$. Celui-ci est composé de l nœuds internes et $(l+m+1)$ points de contrôle. On suppose que tous les nœuds et les points de contrôle sont respectivement de même longueur de description. Les nœuds sont positionnés dans $[x_0, x_{N-1}]$ à des endroits d'observations. Un nœud est alors représenté par un entier entre 0 et $N-1$ et $\log_2(N)$ bits suffisent à son codage.

Quant aux points de contrôle, ce sont des paramètres réels estimés à partir de N données. Lorsque N est grand, la longueur d'un paramètre réel est généralement approchée par $\log_2(\sqrt{N})$. Cependant, chaque point de contrôle ayant une influence locale sur le modèle $g_{\widehat{\theta^l}}$, et le nombre d'observations pouvant être faible, cette approximation ne semble pas adaptée [2]. On considère alors une longueur de description d'un point de contrôle basée sur une loi a priori uniforme sur un intervalle borné $[-\alpha, \alpha]$ [5]. Pour une précision ε sur la description d'un point de contrôle, sa longueur est donnée par $\log_2(\alpha) + 1 - \log_2(\varepsilon)$. Pour $\widehat{\mathbf{t}}$ connu, l'estimation du maximum de vraisemblance pour α est $||\widehat{\mathbf{c}}||_\infty = \max_i |\widehat{c}_i|$. Ainsi, modulo une constante indépendante de l, on obtient le critère MDL

$$\begin{aligned} L(\mathbf{y}) \;\;=\;\; & (m+l+1)\left(\log_2\left(||\widehat{\mathbf{c}}||_\infty\right) + 1 - \log_2(\varepsilon)\right) \\ & + N\log_2\left(RMS\right) + l\log_2(N). \end{aligned} \qquad (6)$$

4.2. Influence de la précision de description de $\widehat{\mathbf{c}}$

Pour $\widehat{\mathbf{t}}$ donné, on étudie l'influence de la précision ε de description de $\widehat{\mathbf{c}}$ sur le modèle reconstruit. Soit $\widetilde{\mathbf{c}}$ une approximation de $\widehat{\mathbf{c}}$ telle que $||\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}||_\infty \leq \varepsilon$. Selon le choix de l'évaluation de l'erreur entre les courbes estimée $\mathbf{B}\widehat{\mathbf{c}}$ et reconstruite $\mathbf{B}\widetilde{\mathbf{c}}$, on détermine une précision ε minimale.

Proposition 4.1 *On a* $||\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}||_\infty \;\leq\; \varepsilon$.

Preuve. D'après (1), on a $||\mathbf{B}||_\infty = 1$ et donc

$$||\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}||_\infty \;\leq\; ||\mathbf{B}||_\infty\,||\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}||_\infty \;\leq\; \varepsilon\,.$$

Ainsi, si l'on souhaite un écart maximal entre les points des courbes $\mathbf{B}\widehat{\mathbf{c}}$ et $\mathbf{B}\widetilde{\mathbf{c}}$ inférieur à l'écart entre les données et $\mathbf{B}\widehat{\mathbf{c}}$, il suffit de fixer $\varepsilon = ||\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}||_\infty$.

Corollaire 4.1 *L'erreur RMS entre les courbes optimale* $\mathbf{B}\widehat{\mathbf{c}}$ *et reconstruite* $\mathbf{B}\widetilde{\mathbf{c}}$ *est inférieure à* ε.

Preuve. On rappelle que pour tout $\mathbf{x} \in \mathbb{R}^N$, on a

$$||\mathbf{x}||_\infty \;\leq\; ||\mathbf{x}||_2 \;\leq\; \sqrt{N}||\mathbf{x}||_\infty \qquad (7)$$

et la RMS entre $\mathbf{B}\widetilde{\mathbf{c}}$ et $\mathbf{B}\widehat{\mathbf{c}}$ vérifie, d'après la prop. 4.1,

$$||\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}||_2 \,/\sqrt{N} \;\leq\; ||\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}||_\infty \;\leq\; \varepsilon\,.$$

Ainsi, pour obtenir une erreur RMS entre les courbes $\mathbf{B}\widehat{\mathbf{c}}$ et $\mathbf{B}\widetilde{\mathbf{c}}$ inférieure à l'erreur RMS entre $\mathbf{y}$ et $\mathbf{B}\widehat{\mathbf{c}}$, il suffit de fixer $\varepsilon = RMS = ||\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}||_2 \,/\sqrt{N}$.

4.3. Critères MDL pour les B-splines

Le critère MDL utilisé est défini par (6). On distingue deux versions selon la précision ε en la supposant fonction de l'erreur de reconstruction du modèle B-spline. On considère qu'il n'est pas nécessaire que l'erreur entre les courbes estimée et reconstruite soit inférieure à l'erreur d'estimation. D'après le paragraphe 4.2, on obtient les versions du critère : (a) : $\varepsilon = RMS$, et (b) : $\varepsilon = ||\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}||_\infty$.

5. PROTOCOLE EXPÉRIMENTAL

L'objectif est d'estimer le modèle B-spline $\widehat{\theta^l}$ et d'établir un compromis, à l'aide des critères MDL ci-dessus, entre sa qualité de modélisation (évaluée par l'erreur RMS) et son nombre de degrés de liberté (d.d.l.) l. On introduit les hypothèses méthodologiques communes à toutes les expériences avant de présenter dans le prochain paragraphe deux expériences permettant d'étudier : la relation entre l'erreur RMS et le nombre de d.d.l. et le comportement des différents critères MDL.

Les expériences sont réalisées sur 500 phrases (soit environ 7000 syllabes) choisies aléatoirement parmi un corpus d'environ 7000 phrases enregistrées. L'enregistrement a été réalisé dans un studio professionnel. Le signal acoustique a été annoté puis segmenté en unités acoustiques. La fréquence laryngienne moyenne, F_0, a été analysée de manière automatique à l'aide de la fonction d'autocorrélation. Ensuite, un algorithme a été appliqué à la chaîne d'unités phonétiques de manière à repérer chaque syllabe.

Pour l nœuds internes, le modèle $\widehat{\theta^l}$ possède $(2l+m+1)$ paramètres et si ce nombre est supérieur au nombre N de valeurs de F_0, il est plus économique de conserver ces dernières. Les courbes ayant un nombre différent d'observations, on normalise le nombre de d.d.l. l pour chaque courbe en le divisant par le nombre N d'observations. Cela est nécessaire pour comparer les moyennes de nombre de d.d.l. Un nombre de d.d.l. normalisé à 1 correspond alors à l'ensemble de ces points.

Toutes les expériences proposées font intervenir le calcul des valeurs moyennes de l'erreur RMS et du nombre de d.d.l. des modèles B-splines. Ces moyennes sont estimées par intervalles de confiance au niveau 99% établis selon une méthodologie bootstrap.

6. EXPÉRIENCES ET RÉSULTATS

Pour chaque courbe observée, le critère MDL sélectionne la structure du modèle parmi l'ensemble des structures de modèles possibles. La première étape consiste à estimer les modèles θ^l en faisant varier l. Ensuite, on applique les critères MDL proposés afin de sélectionner la

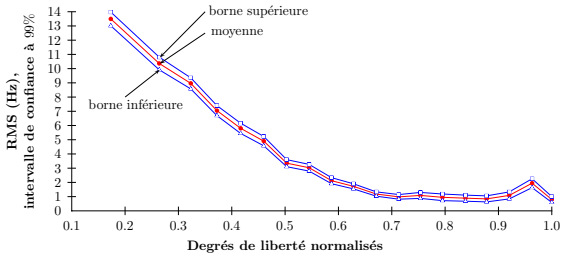


FIG. 1: Évolution des intervalles de confiance à 99% pour l'erreur RMS en fonction du nombre normalisé de d.d.l.

meilleure structure $\widehat{l}$. On présente ci-après les différentes expériences réalisées pour évaluer ces critères.

6.1. Lien entre RMS et nombre de d.d.l.

La figure 1 permet de mesurer l'impact du nombre de paramètres sur la qualité d'estimation. Le nombre normalisé de d.d.l. présenté varie de 0 à 1. Toutes les courbes du corpus sont prises en compte. Lorsque le nombre de d.d.l. augmente, l'erreur RMS diminue. Ce résultat prévisible justifie la recherche d'un compromis entre la précision du modèle et sa complexité. Notamment, on remarque un point d'inflexion dans la courbe pour un nombre de d.d.l. normalisé moyen voisin de 0.65. Cette valeur correspond à une RMS moyenne proche de 1Hz. La remontée de l'erreur moyenne en fin de courbe montre une faiblesse du recuit simulé lorsque le nombre de nœuds devient important. De plus, la condition (4) n'est plus respectée. Bien que l'optimisation par l'algorithme SA soit globale, la solution trouvée n'est pas forcément optimale. Le but du critère MDL étant de déterminer ce compromis, un critère satisfaisant devrait estimer une erreur RMS moyenne et un nombre de d.d.l. moyen aux environs du point d'inflexion de la courbe.

6.2. Analyse des critères MDL proposés

Les résultats de l'évaluation des critères MDL sur le corpus sont résumés dans le tableau 1 (intervalles de confiance à 99%). Deux modes variables ont été testés : l'utilisation de $||\mathbf{y} - \mathbf{B}\hat{\mathbf{c}}||_\infty$ améliore légèrement les résultats de RMS moyenne. En effet, d'après (7), $RMS \leq ||\mathbf{y} - \mathbf{B}\hat{\mathbf{c}}||_\infty$. Le critère (b) est donc moins pénalisant et sélectionne un nombre supérieur de d.d.l. que (a). Si l'on privilégie l'erreur RMS, le critère (b) peut être qualifié de meilleur. De plus, le critère (b) se positionne dans la zone d'inflexion de la courbe au point $(0.627, 0.42)$.

Dans [1] nous avons comparé un modèle B-spline et un modèle spline dans un cadre expérimental assez proche. Nous avions montré, tableau 1 page 4 de l'article, que des modèles B-splines avec un nombre normalisé de d.d.l. de l'ordre de 60% conduisent à une erreur RMS moyenne de l'ordre de 3Hz (on retrouve ce résultat figure 1 abscisse 0.6). Un modèle spline conduisait quant à lui à une erreur RMS moyenne de l'ordre de 12Hz. Dans [6], il a été montré que MoMel conduit à 6Hz d'erreur RMS en moyenne dans le meilleur des cas. Ces résultats suggèrent d'une part qu'un modèle B-spline est plus efficace qu'un modèle spline et d'autre part qu'un critère MDL à précision variable améliore la performance du modèle B-spline.

TAB. 1: Intervalles de confiance à 99% pour l'erreur RMS et le nombre normalisé de d.d.l.

Critères	d.d.l. norm.	RMS (Hz)
(a)	0.599 ± 0.006	0.68 ± 0.11
(b)	0.627 ± 0.006	0.42 ± 0.07

7. CONCLUSION

Dans cet article, nous avons présenté une nouvelle approche pour la modélisation de courbes de F_0 par l'utilisation d'un modèle B-spline. La précision d'une telle modélisation est importante pour caractériser la mélodie en traitement de la parole. Le modèle B-spline généralise le modèle spline et permet de décrire avec précision les irrégularités locales de la courbe. La principale contribution de cet article concerne l'estimation du nombre de paramètres libres du modèle B-spline grâce à une méthodologie MDL. Appliquée à la modélisation syllabique de contours de F_0, cette approche conduit à une erreur RMS moyenne de 0.42Hz pour un nombre moyen de degré de liberté normalisé de 0.63 (un nombre de degré de liberté à 1 signifie qu'on utilise autant de paramètres que de points de la courbe). Ces valeurs de RMS sont d'une part inférieures aux seuils de JND pour le F_0 (de l'ordre de quelques Hz) et d'autre part obtenues avec un facteur de compression relativement important (en moyenne 37%).

RÉFÉRENCES

[1] N. Barbot, O. Boëffard, and D. Lolive. F0 stylisation with a free-knot b-spline model and simulated-annealing optimization. In *Proc. Eurospeech Conf.*, pages 325–328, 2005.

[2] M.A.F. Figueiredo, J.M.N. Leitão, and A.K. Jain. Unsupervised contour representation and estimation using b-splines and a minimum description length criterion. *IEEE Trans. Image Proc.*, 9(6) :1075–1086, 2000.

[3] Mark H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *J. Amer. Stat. Assoc.*, 96(454) :746–774, 2001.

[4] D.J. Hirst, A. Di Cristo, and R. Espesser. Levels of representation and levels of analysis for the description of intonation systems. *In M. Horne (Ed.),Prosody : Theory and Experiment, Kluwer Academic Pusblisher*, 14 :51–87, 2000.

[5] T .C.M. Lee. An introduction to coding theory and the two-part minimum description length principle. *Intl. Stat. Review*, 69(2) :169–183, 2001.

[6] S. Mouline, O. Boëffard, and P.C. Bagshaw. Automatic adaptation of the momel f_0 stylisation algorithm to new corpora. In *Proc. of ICSLP*, 2004.

[7] A. Raux and A.W. Black. A unit selection approach to f0 modeling and its application to emphasis. In *Proc. ASRU Conf.*, pages 700–703, 2003.

[8] S. Sakai and J. Glass. Fundamental frequency modeling for corpus-based speech synthesis based on statistical learning techniques. In *Proc. ASRU Conf.*, pages 712–717, 2003.

[9] P. Taylor. Analysis and synthesis of intonation using the tilt model. *J. Acoust. Soc. America*, 107 :1697–1714, 2000.

Session XVIII

Poster

Jeudi 15 juin 2006 - 16h45 18h00

Constitution d'un corpus textuel basée sur la divergence de Kullback-Leibler pour la synthèse par corpus

Aleksandra Krul[1], *Géraldine Damnati*[1],*Thierry Moudenc*[1], *François Yvon*[2]

[1] France Télécom Division R&D, TECH/SSTP
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{aleksandra.krul,thierry.moudenc,geraldine.damnati}@francetelecom.com

[2] GET/ENST et CNRS/LTCI
46 rue Barrault
75624 Paris Cedex 13
yvon@enst.fr

ABSTRACT

This paper presents a text design method for Text-To-Speech synthesis application. The aim of this method is to build a corpus whose unit distribution is close to a target distribution. As text selection is a NP-hard set covering problem, a greedy algorithm is used. We propose the Kullback-Leibler divergence to compute the score of each candidate sentence. The proposed criterion gives the possibility to control the unit distribution at each step of the algorithm. Finally, we present the first results and we compare the proposed criterion with two standard criteria.

1. INTRODUCTION

La plupart des systèmes de synthèse vocale à partir du texte reposent sur une technique de concaténation d'unités acoustiques pré-enregistrées, la synthèse par corpus étant la plus utilisée. Cette approche repose sur l'utilisation d'une base d'unités acoustiques élémentaires qui résulte de la lecture d'un corpus textuel soigneusement choisi. La qualité du corpus textuel conditionne donc la qualité de la synthèse.

La conception du corpus textuel peut être vue comme un problème de recouvrement d'un ensemble. Chaque phrase du corpus est un ensemble d'unités. L'ensemble cible C contient des unités à couvrir. Le problème consiste à trouver un ensemble de phrases de cardinal minimum dont l'union forme C. Étant donné que le problème est NP-difficile, il n'y a pas d'algorithme exact applicable, d'où le recours à des méthodes heuristiques. L'algorithme glouton est une méthode appropriée pour résoudre ce problème. Il consiste à construire itérativement une solution en ajoutant à chaque pas un élément choisi parmi les autres selon un critère.

Dans le cas de la synthèse de parole, les critères habituellement utilisés découlent de l'objectif principal qui est d'obtenir la couverture des unités. La méthode gloutonne consiste à choisir incrémentalement à partir d'un grand corpus un sous-ensemble de phrases qui atteint la couverture souhaitée. Celle-ci est le pourcentage des unités existantes dans le corpus de départ et présentes dans le corpus construit. Selon les approches, les unités à couvrir sont des diphones, des diphones en contexte, des triphones ou des syllabes. Pour chaque phrase candidate un score est calculé qui permet de choisir la phrase la plus utile, c'est-à-dire celle qui augmente le plus la couverture. La phrase sélectionnée est ensuite retirée du corpus de départ et les unités de celle-ci sont alors enlevées de l'ensemble d'unités à couvrir. De nombreux travaux [8, 2, 6, 3] ont eu recours à l'algorithme glouton pour la constitution du corpus textuel.

D'autres méthodes, inspirées de la méthode gloutonne, ont été proposées notamment la méthode gloutonne inversée (ou cracheuse) [6] et la méthode d'échange par paires [7]. L'algorithme cracheur, à l'inverse de l'algorithme glouton, démarre avec une couverture totale c'est-à-dire celle du corpus de départ. Les phrases sont supprimées une à une jusqu'à ce que la suppression d'une phrase fasse perdre des unités à la couverture totale. Quant à la méthode d'échange par paires, elle vise à améliorer la couverture plutôt qu'à la construire soit en augmentant le nombre d'unités couvertes, soit en diminuant le taux de couverture selon un seuil minimal.

L'efficacité de ces trois méthodes dépend du critère choisi pour calculer le score de chaque phrase candidate. Indépendamment de la méthode utilisée, les critères sont relatifs au nombre d'unités distinctes de la phrase et au nombre d'unités de la couverture. Afin de contrôler la longueur des phrases sélectionnées, le nombre total d'unités dans la phrase est également pris en compte. Dans [6] plusieurs critères ont été présentés et évalués, comme, les critères basés sur le nombre d'unités utiles à la couverture dans les phrases candidates, ou encore la présence d'unités rares dans la phrase.

En fonction de l'objectif à atteindre, le score de chaque phrase candidate peut être calculé de différente manière. Pour atteindre la couverture, le calcul du score le plus simple peut consister à normaliser le nombre d'unités nouvelles d'une phrase par le nombre total d'unités contenues dans cette phrase. Si, en plus d'obtenir une couverture, l'objectif est de favoriser les événements rares alors le calcul du score fait appel aux fréquences des unités observées dans le corpus initial. Dans le but d'obtenir une grande variabilité au niveau phonétique, [3] propose de calculer le score de chaque unité (diphone) de la phrase candidate en fonction de ses contextes phonétiques gauche et droit. Les unités retenues sont celles qui augmentent la variabilité phonétique du corpus. Cette approche permet d'obtenir une meilleure variabilité de triphones dans la base acoustique. L'objectif qui consiste à atteindre la couverture et celui qui vise la variabilité des unités étant partiellement antagonistes, un compromis entre les deux est nécessaire. Cela conduit à des scores différents.

Une des difficultés de la constitution du corpus textuel se trouve dans le choix d'une couverture optimale des unités au sens d'un objectif applicatif. Dans le cas de la synthèse générale, la distribution des unités souhaitée est celle

qui limite la redondance des unités fréquentes et maximise la présence des unités rares. La couverture totale doit être atteinte au moins pour les unités élémentaires. De plus, une représentation suffisante des unités doit être assurée pour anticiper les différents contextes d'apparition possible de ces unités. Pour la synthèse dédiée à des domaines restreints, la distribution idéale des unités est celle qui reflète un contexte applicatif particulier. Dans ce cas, la contrainte sur les unités rares peut être moins forte. En revanche, les unités les plus fréquentes relativement au domaine doivent être bien représentées. La base peut également être plus petite et plus spécifique au domaine visé.

Pour les critères qui cherchent à obtenir une couverture, la distribution des unités dans le corpus final est difficile à maîtriser. C'est pourquoi nous proposons un critère qui vise à contrôler globalement la distribution des unités dans le corpus construit à chaque étape du processus.

Dans cet article nous proposons une méthode gloutonne de constitution de corpus textuel qui repose sur la divergence de Kullback-Leibler. Cette approche vise à construire un corpus dont la distribution des unités tend vers une distribution *a priori*. Le critère utilisé évalue l'utilité d'une phrase en fonction de toutes les unités du corpus construit. Ce critère permet également de maîtriser globalement la distribution des unités dans le corpus. Pour cette étude, la distribution visée est uniforme et l'unité considérée est le diphone. Dans la section 2 nous introduisons la mesure de Kullback-Leibler et nous détaillons notre approche. Enfin, nous présentons les premiers résultats obtenus avec cette méthode et nous la comparons aux méthodes standard qui visent la couverture des unités.

2. APPROCHE ALTERNATIVE

2.1. La mesure de Kullback-Leibler

Avant de détailler notre approche, nous introduisons ici la divergence de Kullback-Leibler (KL). C'est une mesure de similarité entre deux distributions de probabilité P et Q. Elle se calcule de la façon suivante :

$$D(P \parallel Q) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} \qquad (1)$$

La divergence est d'autant plus petite que les distributions sont proches. La divergence de KL est toujours positive, elle est nulle si et seulement si les deux distributions sont identiques [4].

2.2. Sélection de phrases basée sur la divergence de KL

Nous proposons d'utiliser la divergence de KL pour calculer le score d'une phrase dans le processus de sélection de corpus. Le but étant de construire une distribution *a priori*.

Algorithme L'algorithme utilisé est de type glouton. À chaque itération la phrase retenue est celle qui minimise la divergence de KL à la distribution cible. Notons par $S = \{s_1, s_2, \ldots, s_l\}$ le corpus à partir duquel les phrases sont sélectionnées. Nous représentons celles-ci par $S' = \{s'_1, s'_2, \ldots, s'_m\}$, où $m \leq l$.

La distribution *a priori* des probabilités est donnée par Q.

L'estimation des probabilités des unités dans le corpus construit doit se faire sur l'ensemble des unités sélectionnées à chaque itération de l'algorithme. n_i est le nombre d'occurrences de l'unité i dans le corpus candidat qui englobe le corpus déjà construit à l'itération précédente et la phrase candidate. N est la somme du nombre total d'unités déjà sélectionnées et du nombre total d'unités de la phrase candidate. Dans le corpus construit pas à pas la probabilité pour chaque unité est définie par $p_i = \frac{n_i}{N}$, c'est-à dire par sa fréquence d'apparition. Pour les unités qui ne sont pas encore représentées dans le corpus en construction nous considérons que $p_i = 0$. En utilisant la convention $0 \log \frac{0}{q} = 0$, ces unités ne contribuent pas au calcul de la divergence. Le score de chaque phrase est :

$$D(P \parallel Q) = \sum_{i, n_i \neq 0} \frac{n_i}{N} \big(\log \frac{n_i}{N} - \log q_i \big) \qquad (2)$$

À chaque étape, l'algorithme 1 ajoute à l'ensemble de phrases sélectionnées la phrase qui minimise la divergence de KL. La nouvelle distribution des unités ainsi obtenue se rapproche de la distribution visée. L'algorithme s'arrête après avoir inclus au maximum L phrases, où L est une borne fixée à l'avance.

Algorithme 1 Sélection de phrases basée sur la divergence de KL

Définir une distribution cible Q
$S'_0 = \emptyset$
Pour $j = 1$ jusqu'à L **Faire**
  $D_{min} \leftarrow +\infty$
  Pour Chaque phrase $s_k \in S \setminus S'_{j-1}$ **Faire**
    $A_{jk} = S'_{j-1} \cup \{s_k\}$
    Estimer la distribution des probabilités P_{jk} sur l'ensemble A_{jk}
    Calculer $D(P_{jk} \parallel Q)$
    Si $D(P_{jk} \parallel Q) < D_{min}$ **Alors**
      $D_{min} \leftarrow D(P_{jk} \parallel Q)$
      $s_{best} \leftarrow s_k$
    FinSi
  FinPour
  $S'_j \leftarrow S'_{j-1} \cup \{s_{best}\}$
FinPour

Distribution uniforme Pour cette étude l'unité considérée est le diphone. Nous utilisons comme distribution cible la distribution uniforme où toutes les unités sont équiprobables. Cela peut paraître, à première vue, paradoxal. En effet, la distribution des diphones dans le corpus à partir duquel les phrases sont sélectionnées est exponentielle et suit la loi de Zipf [1]. Certaines unités sont très fréquentes, alors que de nombreuses unités sont très peu représentées. En visant la distribution uniforme, le critère tend à mettre au même niveau les unités fréquentes et les unités rares présentes dans le corpus d'origine. Étant donné que l'algorithme sélectionne les phrases entières, la distribution obtenue hérite naturellement de la distribution de départ. Ainsi, nous maximisons la présence des unités rares et nous limitons l'apparition des unités fréquentes dans le corpus final. Ceci se ramène au choix de la distribution d'entropie maximale.

Couverture En visant la distribution uniforme nous introduisons indirectement l'objectif d'atteindre la couver-

ture totale de diphones. En effet, le critère proposé va chercher à prendre toutes les unités distinctes. Cependant, dans la mesure où l'algorithme sélectionne les phrases entières, la distribution obtenue hérite naturellement de la distribution de départ. Nous pouvons ainsi nous attendre à ce que la méthode proposée ne permette pas d'atteindre en peu d'itérations la couverture. Afin d'obtenir la couverture nous avons développé une variante de l'algorithme 1 en ajoutant une contrainte sur l'ensemble des phrases candidates. Tant que la couverture n'est pas atteinte, la phrase retenue est celle qui minimise le critère de KL parmi les phrases qui apportent de nouveaux diphones distincts. L'algorithme reprend son fonctionnement normal une fois la couverture en diphones atteinte.

3. RÉSULTATS EXPÉRIMENTAUX

3.1. Données

Le corpus sur lequel nous avons travaillé contient 7337 phrases issues principalement des articles du journal Le Monde. Il contient également une centaine de phrases utilisées pour des services vocaux. Il a été constitué pour l'enregistrement de la base acoustique. L'objectif de constitution de ce corpus était de couvrir 100% de diphones distincts, 90% de diphones en contextes et 80% de triphones observés dans le corpus général de Le Monde. La longueur maximale de phrases est de 27 mots. Il y a 1170 diphones distincts dans ce corpus. L'intêret d'extraire des phrases de ce corpus est d'observer le comportement des algorithmes dans une optique de réduction de corpus. Dans la mesure où nous n'observons pour l'instant que les diphones distincts, la taille de ce corpus semble suffisante pour cette étude.

3.2. Comparaison des critères

Nous comparons notre critère avec deux critères standard différents. Le premier score est basé sur le nombre de diphones nouveaux présents dans la phrase candidate normalisé par la longueur de la phrase [6]. Le second score est dérivé du premier, mais favorise la sélection des unités rares en pondérant la contribution de chaque unité nouvelle par l'inverse de sa fréquence dans le corpus d'origine [5].

Nous examinons l'évolution de la couverture en fonction de l'itération sur la figure 1 et en fonction du nombre total d'unités sélectionnées sur la figure 2.

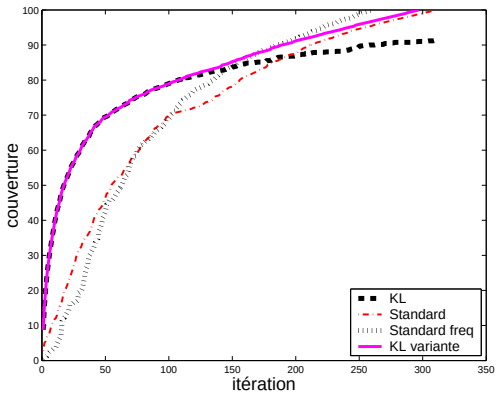


FIG. 1: Couverture

La couverture de diphones est atteinte en peu d'itérations par les méthodes standard et par la variante de notre algorithme ("KL variante"). En revanche, elle n'est atteinte qu'à la fin du processus (soit au bout de 7312 itérations) par l'algorithme 1. Ce phénomène a été observé par [5]. Nous l'expliquons par le fait que la méthode KL préfère sélectionner les phrases qui rendent la distribution des unités dans le corpus plus plate plutôt que les phrases avec des unités nouvelles qui déséquilibrent la distribution. Ainsi, les phrases qui contiennent les unités non-couvertes ne sont pas sélectionnées parce qu'elles ne diminuent pas la divergence de KL. Au début, la méthode proposée choisit les phrases longues qui apportent beaucoup d'unités nouvelles. À la fin du processus, l'algorithme sélectionne des phrases qui perturbent le moins la divergence de KL. Étant donné que le critère reflète de la contribution de chacune des distributions à une mesure globale, la perturbation est moins forte avec des phrases courtes que des phrases longues, même si les phrases courtes ne contiennent pas d'unités nouvelles. Grâce à la modification de l'algorithme la couverture est atteinte en peu d'itérations. Nous observons également,comme nous pouvions nous y attendre, que les couvertures obtenues par notre algorithme et sa variante sont identiques au début du processus.

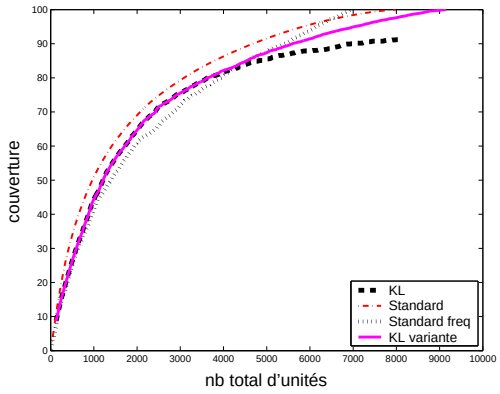


FIG. 2: Couverture

Pour la variante de l'algorithme basé sur la divergence de KL la couverture est atteinte avec le nombre d'unités le plus élevé. Ceci est dû au fait que le critère est indépendant de la longueur des phrases.

Pour examiner le comportement de l'algorithme 1 nous l'avons exécuté jusqu'à ce qu'il n'y ait plus de phrases à choisir : les 7337 ont été sélectionnées. De même, nous avons lancé le processus de sélection de toutes les phrases avec les deux autres algorithmes et calculé ensuite la divergence de KL. La figure 3 illustre cette mesure.

L'allure des courbes de la divergence de KL est similaire pour les trois approches. La distribution des probabilités des unités qui se rapproche le plus de la distribution uniforme est bien celle qui a été construite avec notre méthode. Pour les trois distributions obtenues la divergence de KL diminue rapidement au début du processus, après avoir atteint un minimum elle commence à croître. Toute nouvelle phrase ajoutée à partir de ce minimum augmente la divergence à la distribution uniforme des unités. Ceci et lié au fait que des phrases entières sont sélectionnées et que la distribution des unités tend à suivre la loi de Zipf.

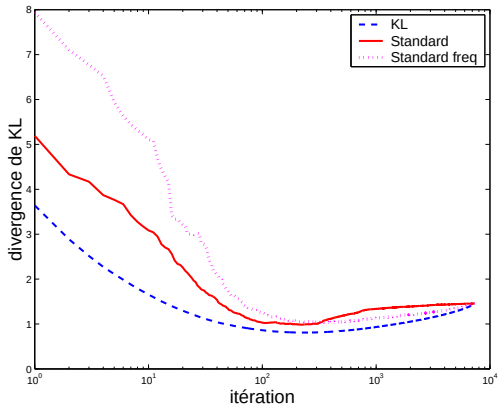


FIG. 3: Divergence de KL

Il est à noter que le minimum est atteint à différentes étapes des algorithmes. Le tableau 1 présente l'état des corpus en construction à l'itération j pour laquelle la divergence de KL est minimale. N_j est le nombre total d'unités à l'itération j.

TAB. 1: État des corpus en construction pour la divergence de KL minimum.

	itération j	N_j	couverture
KL	239	6414	88,46
Standard	218	4852	90
Standard freq	371	10317	100

Pour le critère standard la divergence remonte assez rapidement et ce avec un nombre total d'unités assez faible. Ceci montre que les phrases sélectionnées par ce critère présentent plus de redondance. Le fait de sélectionner en priorité les unités peu fréquentes (Standard freq) retarde la remontée de la courbe. Malgré le fait qu'à ce stade du processus la couverture est atteinte pour cette méthode, la distribution des unités n'est pas celle qui se rapproche le plus de l'uniforme.

3.3. Remarques

Étant donné que, le corpus sur lequel nous avons travaillé résulte déjà d'une sélection, nous avons utilisé un second corpus. Celui-ci contient des phrases choisies aléatoirement dans le corpus général du journal Le Monde. À partir de 20000 phrases, 10263 phrases dont la longueur maximale est de 20 mots ont été retenues. Ce corpus contient 1117 diphones distincts.

De façon générale, le comportement des critères est similaire. Nous observons, par ailleurs, que pour l'obtention de la couverture le nombre d'itérations est plus élevé. De même, la couverture est atteinte avec un nombre total d'unités plus élevé.

4. CONCLUSION ET PERSPECTIVES

Nous avons présenté une méthode alternative pour la construction d'un corpus textuel basée sur la divergence de Kulllback-Leibler. Le critère proposé offre la possibilité de contrôler globalement la distribution des unités dans le corpus final. Son objectif principal est de répartir les unités de façon à ce que leur distribution se rapproche d'une distribution cible. Pour cette étude la distribution uniforme des unités est visée. Des distributions adaptées à des domaines spécifiques doivent être envisagées. L'avantage de cette méthode est qu'elle permet de viser différentes distributions. L'adaptation du corpus textuel à un contexte applicatif peut être facilement réalisée.

De plus, pour atteindre la couverture des unités nous avons implémenté une variante de l'algorithme. Grâce à la modification de l'algorithme la couverture est atteinte et la distribution des unités reste la plus proche de la distribution souhaitée.

Toutefois, pour valider la méthode proposée, des tests sur des corpus de taille plus importante doivent être effectués. Nous envisageons également de travailler sur d'autres types d'unités, par exemple sur des diphones en contexte ou encore sur des triphones.

Enfin, des évaluations des critères sont envisagées d'un point de vue applicatif en comparant la qualité de la synthèse obtenue avec ces critères.

RÉFÉRENCES

[1] R.H. Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

[2] A. W. Black and K. A. Lenzo. Optimal Data Selection for Unit Selection Synthesis. In 4^{rd} *ESCA Workshop on Speech Synthesis*, Scotland, 2001.

[3] B. Boozkurt, O. Ozturk, and T. Dutoit. Text design for TTS speech corpus building using a modified greedy selection. In 8^{th} *European Conference on Speech Communication and Technology (Eurospeech)*, pages 277–280, Geneva, Switzerland, September 2003.

[4] T.M Cover and J.A Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[5] Y. Feng. Selection of text script for text-to-speech synthsesis. In 5^{th} *IASTED International Conference SIGNAL AND IMAGE PROCESSSING*, Honolulu, Hawaii, USA, August 2003.

[6] H. François. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. PhD thesis, Université de Rennes 1, 2002.

[7] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu. A Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody. In 6^{th} *International Conference on Spoken Language Processing (ICSLP)*, pages 277–280, Beijing, China, September 2000.

[8] J.P.H van Santen and A. L. Buchsbaum. Methods for Optimal Text Selection. In 5^{th} *European Conference on Speech Communication and Technology (Eurospeech)*, pages 553–556, Rhodes, Greece, September 1997.

eLite : système de synthèse de la parole à orientation linguistique

Richard Beaufort, Alain Ruelle

Multitel ASBL
Avenue Copernic 1, 7000 Mons, Belgique
{beaufort,ruelle}@multitel.be
http://www.multitel.be/TTS

ABSTRACT

eLite is the Text-to-Speech synthesis system developed by the TTS-NLP group of Multitel ASBL. The creation of eLite has been an opportunity for the group to carry out and integrate further research on all domains of text-to-speech synthesis, like morphological analysis, syntactic desambiguation and non-uniform units selection. This paper presents the general features and techniques of the system.

1. INTRODUCTION

En synthèse de la parole à partir du texte, la génération du signal de parole n'est pas directement réalisée à partir du texte, mais à partir d'une représentation phonétique et prosodique de celui-ci. Cependant, parce que la langue écrite regorge d'ambiguïtés linguistiques, la génération de la représentation phonétique et prosodique doit elle-même être précédée d'une phase de désambiguïsation du texte. C'est conscient de l'importance d'une analyse linguistique fiable que le groupe TTS-NLP de Multitel ASBL a développé eLite, prononcé [i l a j t], dont le nom signifie ***E**nhanced, **LI**nguistically-based **TE**xt-to-speech synthesis system*.
En ce qui concerne l'étape de génération du signal de parole, les premières versions d'eLite ont intégré le synthétiseur MBROLA, basé sur le principe de concaténation d'unités de parole pré-enregistrées et prosodiquement neutres. La dernière version d'eLite intègre le dernier état de l'art en synthèse, la sélection et la concaténation d'unités non-uniformes : la sélection est réalisée par l'algorithme LiONS, la concaténation, par TP-MBROLA.

Commencé en septembre 2001 et conçu initialement afin de fournir à l'équipe une plateforme complète de test pour de nouveaux algorithmes utiles à la synthèse, eLite est rapidement devenu un logiciel stable, robuste et rapide, dont des versions de démonstration sont disponibles sur le site du groupe (http://www.multitel.be/TTS).

Cet article présente l'architecture générale d'eLite ainsi que les différents modules du processus de synthèse.

2. ARCHITECTURE DU SYSTÈME

Unité linguistique. La totalité de l'architecture d'eLite repose sur une notion fondamentale, celle d'*unité linguistique*. Une unité linguistique, dans eLite, est *un mot ou une séquence de mots et de symboles formant un tout*. L'unité linguistique de base est évidemment le mot, dans le sens de *séquence de caractères alphabétiques comprise entre deux espaces*, l'espace pouvant être un ou plusieurs blancs, retours à la ligne ou signes de ponctuation. Une unité linguistique peut également correspondre à un mot composé dont les constituants sont séparés par un ou plusieurs blancs ou par un tiret. Enfin, l'unité linguistique peut être une *unité de sens*, comme les adresses URL, les numéros de téléphone ou les nombres et unités de mesure.

Modules et Structure de données. eLite (cf. fig. 1) reçoit en entrée un texte et produit en sortie la parole correspondante. Le système se divise en 3 modules principaux :

1. Le NLP (*Natural Language Processing*), qui gère le traitement du langage naturel.
2. La Sélection, qui choisit les unités de parole.
3. Le DSP (*Digital Signal Processing*), qui concatène les unités de parole choisies et produit le signal voulu.

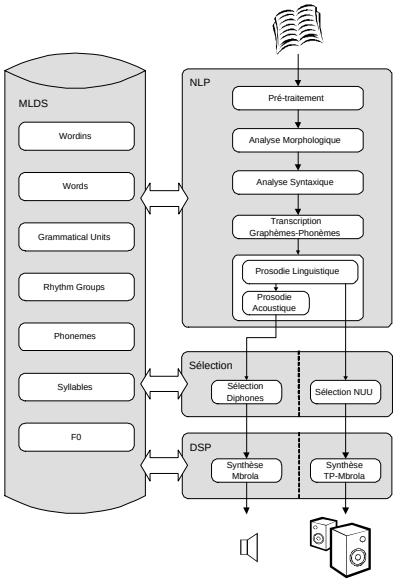


FIG. 1: Architecture d'eLite

Le module NLP se divise lui-même en 5 étapes. Le pré-traitement, l'analyse morphologique et l'analyse syntaxique désambiguïsent le texte. La conversion graphèmes-phonèmes associe une séquence de phonèmes à chaque mot, et gère les phénomènes phonétiques aux frontières des mots. La génération de la prosodie génère des informations prosodiques de type linguistique et

acoustique. Notons que la prosodie acoustique n'est nécessaire que dans le cas d'une synthèse via MBROLA.

Les 3 modules communiquent au travers d'une structure de données multicouches, la MLDS (*Multi-Layers Data Structure*), inspirée de celle proposée par le projet Festival de l'université d'Edinbourg [3]. Notre MLDS se compose de 7 couches :

1. *Wordin* : ce sont les unités linguistiques détectables par le pré-traitement, telles que les URLs, téléphones, etc. Notons qu'un mot composé ne sera par détecté comme un seul Wordin, mais comme plusieurs.
2. *Word* : ce sont les mots au sens défini précédemment. Un Word peut donc être une partie de Wordin, ou correspondre à un Wordin complet. Chaque Word possède une liste de natures possibles.
3. *Grammatical Unit* : une Grammatical Unit est l'analyse syntaxique d'une unité linguistique. Une Grammatical Unit correspond donc généralement à un Wordin. Cependant, dans le cas des mots composés, une Grammatical Unit correspondra à plusieurs Wordins.
4. *Rhythm Group* : ce sont les groupes de souffle d'une phrase, entre lesquels une pause peut survenir. Un Rhythm Group englobe généralement plusieurs Grammatical Units.
5. *Phoneme* : les phonèmes des mots du texte. Un phonème est toujours lié à un Word.
6. *Syllable* : les syllabes, constituées de phonèmes.
7. *F0* : les fréquences fondamentales attribuées aux phonèmes voisés.

Niveaux d'analyse. L'originalité d'eLite est qu'il présente deux niveaux d'analyse syntaxique, grâce à la présence simultanée d'informations syntaxiques dans la couche *Grammatical Unit* et dans les natures de la couche *Word* (Cf. tab. 1).
La couche *Grammatical Unit* donne une vue macroscopique de la phrase, puisqu'elle ne propose une analyse syntaxique que pour les unités linguistiques. Ceci sera particulièrement utile lors de l'analyse syntaxique : de la sorte, la syntaxe n'a pas à s'embarrasser de détails comme, par exemple, de la présence de symboles dans une URL.
Les listes de natures de la couche *Word* fournissent une analyse détaillée de chaque mot. Ceci sera entre autres nécessaire lors de la phase de conversion graphèmes-phonèmes. Il sera par exemple très utile de savoir que l'URL \textit{www.president.fr} contient le nom *président* et non le verbe, afin de générer la phonétisation correcte.

TAB. 1: Words, Natures, Grammatical Units

Word	Natures	Gram. Unit
pommes	NOUN	NOUN
de	PREP	
terre	NOUN	

Langues. Actuellement, eLite traite le français et l'anglais. Néanmoins, le système est ouvert à d'autres langues. En effet, toutes les données nécessaires aux différents processus ont été externalisées. L'ajout d'une langue dans eLite revient donc à concevoir les bases de données qui y correspondent. eLite est cependant actuellement limité aux langues dites *romanes* (français, espagnol, italien, etc.) et *germaniques* (anglais, allemand, néerlandais, etc.), parce que les étapes de l'analyse restent intrinsèquement liées à la structure de ces langues. Des langues telles que le turc, l'arabe ou le chinois ne sont donc pas encore gérables dans eLite.

3. NLP

Le but du NLP est de fournir aux modules suivants une représentation phonétique et prosodique du texte à synthétiser. Toutefois, les ambiguïtés présentes dans la langue écrite nécessitent de commencer par désambiguïser le texte, ce qui est réalisé en 3 étapes : pré-traitement, analyse morphologique et analyse syntaxique.

Pré-traitement. Le rôle du pré-processeur est de diviser le texte en Wordins et de supprimer les caractères parasites (espaces, caractères vides de sens). Les Wordins générés correspondent à des unités linguistiques, sauf dans le cas des mots composés (Cf. *Analyse morphologique*).
Les unités linguistiques à détecter sont modélisées à l'aide d'expressions régulières compilées sous la forme d'une machine à états finis, chargée par le pré-processeur. Les unités linguistiques reconnues sont :

– *Mot* : mangera, ils, TCTS
– *Ponctuation* : . ! ? ; : - . . .
– *URL* : www.cuisiner.fr/index.php?x=10, 10.108.55.9, john.smith@foo.co.uk
– *Date* : 01/02/2002, 30/06/75
– *Heure* : 8h 10min 30s, 10:45
– *Téléphone* : 010/24.38.97, +33 3 27 33 34 54
– *Nombres* : 10.000.000, -10.045,43e-43
– *Montants* : $40000, -10.000 EUR, +43,433.76 USD
– *Mesures* : 25 km/h, 10.343,45 N/m^2
– *Acronymes* : A.S.B.L., T.C.T.S., O.T.A.N.

Lorsqu'un Wordin est détecté, le pré-processeur le segmente en Word à l'aide d'une autre machine à états finis, également générée à partir d'expressions régulières. Par exemple, une URL sera segmentée en symboles (:, ., //), acronymes (http, ftp, www, etc.) et mots (Cf. tab. 2).
Le pré-processeur effectue également une détection sommaire de la mise en page du document. Celle-ci se borne pour l'instant à repérer les titres (chiffres numériques ou romains suivis d'un point en début de ligne), les énumérations (points, astérisques ou tirets en début de ligne) et les paragaphes (multiples sauts de lignes), ce qui permet d'identifier des fins de phrase non marquées par un symbole de ponctuation et d'insérer des pauses de longueur adaptée dans la prononciation du texte.

TAB. 2: Segmentation Wordin → Words

Wordins	Words	
http://www.fortis.be	http	ACRONYM
	:	SYMBOL
	//	SYMBOL
	www	ACRONYM
	.	SYMBOL
	fortis	NOUN
	.	SYMBOL
	be	ACRONYM

Analyse morphologique. Cette analyse s'effectue en deux étapes. Dans un premier temps, elle attribue à chaque Word une liste de natures possibles. Dans un second temps, elle constitue la couche Grammatical Unit.

Pour attribuer une liste de natures à un Word donné, l'analyseur applique les règles suivantes dans l'ordre, et s'interrompt dès qu'une règle produit au moins une analyse :

1. *Wordins spéciaux* :les Wordins qui ne sont pas de type « Mot » peuvent présenter des caractères spéciaux qui seront analysés différemment en fonction du type de Wordin. Par exemple, un point (.) sera considéré comme un symbole s'il appartient à une URL ou à un Téléphone, mais sera analysé comme une ponctuation forte si le Wordin est de type Ponctuation.
2. *Test flexionnel* : cette analyse est tout-à-fait classique. L'analyseur cherche dans le mot des flexions potentielles, et les remplace par les flexions normalisées correspondantes. Si la forme normalisée appartient au dictionnaire de lemme, l'analyse est conservée.
3. *Test de réaccentuation* : l'analyseur cherche les caractères qui pourraient avoir été désaccentués. Par exemple, le mot *eleve* peut être la forme désaccentuée de *élevé* ou *élève*. Chaque forme réaccentuée subit le traitement décrit au point 2. Ceci est particulièrement pertinent pour les URLs, toujours désaccentuées.
4. *Catégories ouvertes* : en l'absence de correcteur orthographique, l'analyseur doit conclure que le mot inconnu est probablement un néologisme. Pour cette raison, il lui attribue la liste des catégories ouvertes aux néologismes. En français, il s'agit de *nom*, *verbe*, *adjectif* et *adverbe*.

Généralement, la structure de la couche Grammatical Unit correspond à celle de la couche Wordin, puisqu'un Wordin est le plus souvent une unité linguistique complète. Cependant, les mots composés sont séparés en plusieurs Wordins par le pré-traitement. De ce fait, l'analyseur effectue une étape de recomposition de manière à regrouper plusieurs Wordins en une seule Grammatical Unit. La recomposition se calcule de proche en proche, en fonction de règles décrites dans un fichier. Un exemple de règle : *préfixe + tiret + infinitif* $\rightarrow$ *infinitif*. Cette régle s'applique par exemple à *sur-couver*, absent du dictionnaire mais recomposé par l'analyse.

Analyse syntaxique. L'analyse consiste, pour une suite de mots $\{w_1, w_2, \ldots, w_n\} = W$, à déterminer la meilleure suite de catégories $\{t_1, t_2, \ldots, t_n\} = T_{MAX}$. Par la règle de Bayes, ceci se formalise :

$$T_{MAX} = \arg\max_T P(W|T)P(T) \quad (1)$$

Le modèle de langue, $P(T)$ est classiquement réduit à un modèle *n*-gramme lissé. Le lissage utilisé ici est une *interpolation linéaire*. Le modèle de mots $P(W|T)$, classiquement ignoré parce que difficile à estimer, est réintroduit sous la forme d'un modèle de classes d'ambiguïté lexicales $P(A|T)$. Nos tests, réalisés à partir du corpus d'entraînement par *10-fold-cross-validation*, montrent que le système d'analyse profite à la fois de l'interpolation linéaire sur le modèle de langue, et de la réintroduction du modèle de mots au travers d'un modèle de classes. En moyenne, le système donne 94,5% d'étiquetage grammatical correct. Un article complet est dédié à notre analyse syntaxique [2].

Conversion graphèmes-phonèmes. Le module produit la transcription phonétique des mots du texte. Pour chaque graphème d'un mot considéré hors contexte, le système décide, à l'aide d'un arbre de décision compilé à partir d'un dictionnaire d'apprentissage [9], du phonème qui lui correspond en tenant compte de son contexte graphémique et de la nature lexicale du mot.
Le mot est ensuite remis en contexte, où les phénomènes phonétiques traités sont :
– la liaison : les oiseaux $\rightarrow$ [l e z w a z o]
– l'amuïssement : cinq mille $\rightarrow$ [s ɛ̃ _ m i l]
– la lubrification du discours : quelques patients $\rightarrow$ [k ɛ l k ə p a s j ɑ̃], quelques amis $\rightarrow$ [k ɛ l k ə z a m i]

Génération de la prosodie. Le module se divise en 2 parties : prosodie linguistique et prosodie acoustique. La prosodie linguistique génère les groupes de souffle (*Rhythm Group*) et la syllabation des phrases, à l'aide d'algorithmes de type chink/chunk. Ces informations sont suffisantes pour la synthèse par sélection d'unités non-uniformes via LiONS [6], mais incomplètes pour la synthèse par concaténation de diphones via MBROLA.
La prosodie acoustique, dans eLite, est basée sur un arbre de régression et de classification (CARTs), entraîné sur un corpus de parole qui, pour chaque phonème du corpus, recense le nombre de phonèmes et l'accent symbolique de la syllabe à laquelle il appartient [8]. Pour un phonème donné, le CARTs détermine sa durée, mais également sa fréquence fondamentale s'il s'agit d'un phonème voisé.

4. SÉLECTION ET DSP

MBROLA fait partie des premiers systèmes de synthèse à utiliser une base de données vocales pour la génération du signal de parole. Dans ces systèmes, un seul exemplaire de chaque unité de parole (généralement des diphones) est représenté dans la base. L'idée sous-jacente est de regénérer les informations prosodiques (F0, durée) au moment de la synthèse dans le DSP. Malheureusement, ce traitement du signal entraîne une détérioration de la qualité et du naturel de la voix de synthèse.
Pour conférer à la synthèse un caractère plus naturel, voire proche de la parole humaine, les chercheurs ont voulu mettre en œuvre le principe de *choose the best to modify the least* [1] : au lieu de ne contenir qu'un exemplaire des unités de parole, le corpus employé en compte plusieurs, non neutralisés et donc prosodiquement variables. Cette variabilité dans la base a permis à la phase de sélection de rechercher les unités de parole qui correpondent au mieux aux unités décrites par le NLP (coût cible), et qui se concatènent au mieux (coût de concaténation) de manière à éviter autant que possible les modifications du signal. C'est cette méthode de recherche, basée sur ce double coût, qui a donné naissance à la notion d'*unités non-uniformes*. LiONS et TP-MBROLA appartiennent à ce dernier état de l'art.

MBROLA. La sélection qui précède MBROLA est minimale. Elle consiste simplement à générer un fichier dans lequel chaque ligne décrit un phonème : son nom, sa durée et, si le phonème est voisé, l'évolution de sa fréquence fondamentale.
MBROLA (*Multi-Band Re-synthesis OverLap Add*) travaille en deux phases [7]. Dans un premier temps, il extrait d'une base de données les unités acoustiques décrites dans le fichier de la sélection, et leur applique la prosodie

demandée. Dans un second temps, ces unités sont concaténées, et un lissage spectral est entrepris aux frontières des unités, de manière à éviter toute discontinuité acoustique.

LiONS et TP-MBROLA. LiONS réalise la phase de sélection des unités non-uniformes [6]. L'originalité de ce système est qu'il ne sélectionne pas les unités de parole sur la base de critères acoustiques (F0, durée, tons), mais sur la base d'informations linguistiques uniquement : position du phonème dans la syllabe, de la syllabe dans le mot et dans le groupe rythmique, contexte articulatoire, etc. Cette approche, qui donne une plus grande liberté à la courbe mélodique en évitant de répéter à l'infini les mêmes patrons prosodiques de phrase en phrase, donne à la voix de synthèse un plus grand naturel.
Le synthétiseur TP-MBROLA (*True Period MBROLA*) [4] se contente d'extraire de la base de données les unités choisies par la sélection, et les concatène en appliquant un lissage de type Overlap and Add [5] uniquement aux frontières des unités.

5. Evaluation du NLP

Le NLP a été évalué sous Windows, sur une architecture Intel Pentium Mobile 1.7 GHz pourvue d'1 Go de RAM. Les bases de données du NLP, non optimisées, représentent 3,5 Mo sur le disque et 35 Mo en RAM pour le français, 4 Mo sur le disque et 36 Mo en RAM pour l'anglais. Leur chargement en RAM prend 1,06 sec.
Les performances du NLP ont été évaluées sur un texte contenant 69.033 mots (427.426 caractères), soit environ 8 heures de parole. Sur ce corpus, le temps de traitement est de 93,684 sec, ce qui représente : 4.562,423 caractères/sec ou 736,871 mots/sec. En termes de durée de parole générée par seconde, le NLP est donc environ 306 fois temps réel.

6. Démonstrateurs

Trois démonstrations figurent sur www.multitel.be/TTS :
– *eLite Demo*. Cette application téléchargeable est disponible en français et en anglais, et fonctionne sous Windows et Linux. eLite Demo intègre uniquement MBROLA, et inclut : les bases de données nécessaires, une interface graphique de test et une bibliothèque dynamique qui peut être intégrée dans un programme tiers. Il s'agit d'une version ralentie d'eLite, qui ne peut être employée à des fins commerciales ni militaires.
– *eLite OnLine*. Il s'agit d'une interface de démonstration en ligne intégrant également MBROLA. L'utilisateur dispose ici d'une palette plus vaste de voix de synthèse (7 pour le français et 4 pour l'anglais), et peut obtenir les résultats d'analyse du texte module par module.
– *LiONS*. Des échantillons de la synthèse obtenue à partir de LiONS peuvent être écoutés sur notre site.

7. Conclusion et perspectives

eLite est un système complet de synthèse de la parole à partir du texte, qui accorde une importance certaine à l'analyse linguistique, étape-clef du processus de synthèse. Initialement dédié au synthétiseur MBROLA, eLite intègre dans sa dernière version une synthèse par sélection d'unités non-uniformes, via LiONS et TP-MBROLA.

Les perspectives d'évolution d'eLite sont importantes. Les sujets d'étude du groupe sont actuellement :
– *La correction orthographique*. Le problème en synthèse se situe au niveau des fautes audibles comme l'absence d'un accord ou une mauvaise épellation.
– *La détection de mise en page*. Un document est un ensemble de blocs : paragraphes, tables, adresses, signatures, colonnes. Détecter ces blocs permettra d'éviter une lecture linéaire et absurde du document.
– *La synthèse émotionelle*. L'idée est d'insérer de l'émotion (colère, joie, tristesse, etc.) dans la parole de synthèse obtenue par sélection d'unités non-uniformes.

8. Remerciements

Nos plus vifs remerciements vont à Vincent Pagel, qui a lancé eLite et a développé la première version du phonétiseur, et à Xavier Ricco, auteur de la première version de l'analyseur morphologique. Nous tenons à remercier tout particulièrement Fabrice Malfrère, qui a développé le générateur de prosodie acoustique, Vincent Colotte, qui a travaillé à la mise au point de LiONS, et Baris Bozkurt, auteur de TP-MBROLA. Nous remercions également Thierry Dutoit, dont les conseils avisés guident nos recherches. Enfin, nous adressons notre sincère reconnaissance à Piet Mertens, Dominique Wynsberghe et Michel Bagein, qui ont eu la patience de récolter et de construire les bases de données de l'analyseur morphologique.

Références

[1] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri. Choose the best to modify the least : A new generation concatenative synthesis system. In *Proceedings of Eurospeech'95*, volume 1, pages 581–584, Madrid, Spain, 1999.

[2] R. Beaufort, T. Dutoit, and V. Pagel. Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïtés lexicales. In *Actes des JEP*, pages 133–136, Nancy, France, 2002.

[3] A.W. Black, P. Taylor, and R. Caley. *The Festival Speech Synthesis System : System Documentation*. University of Edinburgh, 1997.

[4] B. Bozkurt, C. d'Alessandro, T. Dutoit, V. Pagel, and R. Prudon. Improving Quality of MBROLA Synthesis for Non-Uniform Units Synthesis. In *Proceedings of the IEEE TTS 2002 Workshop*, 2002.

[5] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of ICASSP'86*, pages 2015–2018, Tokyo, Japan, 1986.

[6] V. Colotte and R. Beaufort. Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes : LiONS. In *Actes des JEP*, Fès, Maroc, 2004.

[7] T. Dutoit and V. Pagel. Le projet MBROLA : vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. In *Actes des JEP*, pages 441–444, Avignon, France, 1996.

[8] F. Malfrere, T. Dutoit, and P. Mertens. Fully Automatic Prosody Generator for Text-to-Speech Synthesis. In *Proceedings of ICSLP*, pages 1395–1398, Sydney, Australia, 1998.

[9] V. Pagel, K. Lenzo, and A. W. Black. Letter-to-Sound Rules for Accented Lexicon Compression. In *Proceedings of ICSLP*, pages 252–255, Sydney, Australia, 1998.

Coopération entre méthodes locales et globales pour la segmentation automatique de corpus dédiés à la synthèse vocale *

Safaa Jarifi[1], *Olivier Rosec*[2], *Dominique Pastor*[1]

[1] ENST Bretagne, 29285 Brest Cedex, France
{safaa.jarifi,dominique.pastor}@enst-bretagne.fr

[2] France Télécom, Division R&D TECH/SSTP/VMI,
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
olivier.rosec@francetelecom.com

ABSTRACT

This paper introduces a new approach for the automatic segmentation of corpora dedicated to speech synthesis. The main idea behind this approach is to merge the outputs of three segmentation algorithms. The first one is the standard HMM-based (Hidden Markov Model) approach. The second algorithm uses a phone boundaries model, namely a GMM (Gaussian Mixture Model). The third method is based on Brandt's GLR (Generalized Likelihood Ratio) and aims to detect signal discontinuities in the vicinity of the HMM boundaries. Different fusion strategies are considered for each phonetic class. The experiments presented in this paper show that the proposed approach yields better accuracy than existing methods.

1. Introduction

L'approche de synthèse par corpus (SPC) repose sur la concaténation de segments de parole contenus dans une grande base de données enregistrée par un locuteur professionnel. Le succès de cette technologie tient au fait que, moyennant une couverture acoustico-prosodique suffisante, il devient possible de sélectionner une séquence d'unités acoustiques correspondant au contexte de synthèse. De ce fait, les modifications des unités de synthèse peuvent être sinon évitées, du moins limitées, ce qui permet de préserver le naturel de la parole synthétique ainsi produite. Cependant, avec la SPC, la création de nouvelles voix de synthèse devient extrêmement coûteuse, car, outre l'enregistrement du corpus proprement dit, de nombreux traitements doivent être effectuées pour obtenir un dictionnaire acoustique utilisable par un système de synthèse. Parmi ceux-ci, les tâches de phonétisation mais surtout de segmentation du corpus sont particulièrement critiques. En effet, même lorsque la chaîne phonétique correspondant à l'énoncé enregistré est connue, les méthodes de segmentation automatiques actuelles sont jugées trop peu précises pour pouvoir être utilisées telles quelles dans le processus de création de voix. Par conséquent, une étape de vérification manuelle de la segmentation demeure nécessaire. Cette étape, de loin la plus coûteuse, est un véritable frein à la diversification de voix dans le cadre de la SPC.

L'automatisation du processus de segmentation de la parole revêt donc une importance particulière pour la diversification de voix dans le cadre de la SPC. L'approche la plus répandue en segmentation de la parole et offrant les meilleurs résultats est celle reposant sur l'utilisation de modèles HMM [5]. Cette méthode peut être considérée comme contrainte linguistiquement car elle prend en entrée la chaîne phonétique, supposée exacte et obtenue par étiquetage manuel, correspondant à l'énoncé enregistré pour en déduire une séquence de modèles HMM. Elle consiste alors à effectuer un alignement forcé de cette séquence de modèles HMM sur le signal de parole. La principale limite de cette approche tient au fait que les HMM sont surtout réputés pour leur capacité à modéliser les zones stables des phones et non pas à détecter de manière fine des ruptures dans le signal de parole. D'autres approches telles que l'algorithme de Brandt [6] ont également été proposées pour localiser des ruptures dans le signal de parole. Ces méthodes sont a priori assez bien adaptées pour une tâche de segmentation, mais, n'étant pas contraintes sur le plan linguistique, elles produisent des omissions et des insertions de marques de segmentation.

Dans cet article, nous combinons ces deux types de méthodes. Dans cette optique, trois algorithmes de segmentation sont utilisés. Le premier est un algorithme classique de segmentation par HMM. Le deuxième est un algorithme d'ajustement des marques de segmentation par le biais d'une modélisation des frontières de phones par GMM [3]. Le troisième est une version modifiée de l'algorithme de Brandt de manière à rendre les marques de segmentation produites par cet algorithme compatibles avec la séquence phonétique réalisée. Ces différents algorithmes sont décrits en section 2. En section 3, nous présentons les différentes stratégies de combinaison envisagées pour ces algorithmes. Une expérimentation sur un corpus de parole dédiée à la synthèse vocale est également réalisée pour valider la méthode proposée.

2. Méthodes de segmentation

2.1. Segmentation par HMM

Les approches par HMM sont actuellement considérées comme un standard dans le domaine de la segmentation de la parole. Leur mise en œuvre requiert deux étapes : d'une part une phase d'apprentissage visant à estimer les modèles acoustiques et d'autre part une phase d'application de ces modèles à des fins de segmentation. L'étape de segmentation

*Cette étude est soutenue par France Télécom.

proprement dite revient à utiliser l'algorithme de Viterbi pour effectuer un alignement forcé entre la séquence de HMM correspondant à la séquence phonétique d'entrée et le signal de parole.

La phase d'apprentissage est cruciale, car la précision d'une segmentation est étroitement liée à la qualité de l'estimation des modèles. Nous utilisons ici une méthode classique basée sur une estimation initiale des modèles acoustiques par l'algorithme de Baum-Welch et suivie d'une procédure itérative au cours de laquelle les modèles sont réactualisés sur la base d'un alignement forcé [2]. Cette méthode d'apprentissage sera appliquée à l'ensemble du corpus de parole et dénommée dans la suite *HMM1*. En outre, nous nous proposons également d'étudier les performances d'un système de segmentation par HMM, lorsque les modèles acoustiques sont initialisés sur une partie de la base d'apprentissage pour laquelle une segmentation manuelle est disponible. Ces modèles sont ensuite utilisés pour segmenter tout le corpus. Une telle stratégie appelée par la suite *HMM2* offre généralement de meilleurs résultats puisque l'initialisation des modèles acoustiques est *a priori* meilleure [4].

2.2. Post-traitement par modèle de frontière

Dans [3], Wang et al. adjoignent à l'algorithme de segmentation par HMM un post-traitement effectué au voisinage des frontières de phones et utilisant des modèles GMM. Cette méthode de segmentation requiert une estimation préalable des modèles de frontières à partir d'un petit corpus segmenté manuellement. Plus précisément, étant donné une marque de segmentation manuelle, un super-vecteur est construit par la concaténation des N vecteurs acoustiques de taille N_c associés aux N trames de part et d'autre de la trame contenant la frontière, conformément à la figure 1. Lors de cette phase d'apprentissage supervisée, une classification des frontières par arbre de décision est opérée et pour chacune des classes obtenues, un modèle GMM est estimé. Une fois le modèle de frontière appris, le processus de correction consiste à déterminer, au voisinage de la marque de segmentation obtenue par alignement forcé, l'instant pour lequel la vraisemblance du super-vecteur par rapport au modèle de frontière considéré est maximale.

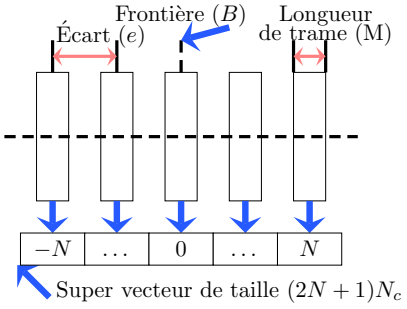


Fig. 1: Constitution d'un super-vecteur

2.3. Algorithme de Brandt

L'algorithme de Brandt [6] permet de détecter des ruptures de stationnarité dans un signal de parole. Il fait l'hypothèse que le signal de parole $y(n)$ est une suite de segments stationnaires et que le signal sur chacun de ces segments est modélisable par un modèle autorégressif (AR) : $y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n$, où p est l'ordre du modèle supposé constant pour tous les segments et où e_n est un bruit blanc gaussien de moyenne nulle et de variance σ^2. Par conséquent, chaque unité est associée à un vecteur de paramètres $\Theta = (a_1 = \ldots, a_p, \sigma)$.

Soit w_0 une fenêtre d'observation de longueur M. Le principe de l'algorithme de Brandt est de décider si w_0 doit être découpé en deux fenêtres w_1 et w_2 ou non. Cette décision se fait sur les vecteurs de paramètres Θ_1 et Θ_2 associés respectivement aux fenêtres w_1 et w_2. Ainsi, un changement entre Θ_1 et Θ_2 est détecté quand le rapport de vraisemblance généralisé (GLR) dépasse un seuil λ prédéfini. L'instant de ce changement est considéré comme l'instant de rupture de stationnarité. Comme mentionné précédemment, cet algorithme n'est pas contraint linguistiquement et par conséquent engendre des omissions et des insertions de marques de segmentation, dont les taux varient en fonction du seuil de détection choisi.

Dans le cadre de la segmentation de corpus dédiés à la synthèse vocale, nous disposons de l'information de phonétisation, voire d'une segmentation en accord avec cette séquence phonétique, obtenue par exemple via une méthode de type HMM. Pour une chaîne phonétique de taille L commençant et se terminant par un silence, notons $U = (U_0, U_1, \ldots, U_L)$ la séquence des marques de segmentation ainsi déterminée. A partir de ces marques de segmentation initiales, nous pouvons alors définir des intervalles temporels sur lesquels sont susceptibles de se produire les transitions entre les différents phones de la séquence phonétique. Ces intervalles sont de la forme $I_i = [V_i, V_{i+1}]$ avec $V_i = \frac{(U_{i-1}+U_i)}{2}$ pour tout i dans $\{1, \ldots, L-1\}$. Nous appliquons alors un algorithme de Brandt modifié sur chacun des intervalles I_i : la détection de rupture par seuillage du GLR est ici remplacée par la maximisation du GLR sur I_i, ce qui permet d'éviter tout risque d'insertion et d'omission.

2.4. Évaluation des algorithmes

Dans cette section, nous évaluons les différents algorithmes présentés ci-dessus sur un corpus de parole dédié à la synthèse vocale pour le français. Ce corpus comporte 7300 phrases prononcées par un sujet féminin et échantillonnées à 16 kHz. L'analyse acoustique et l'apprentissage des HMM sont effectués via l'outil HTK [1]. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des lois multigaussiennes. Le nombre de gaussiennes est fixé à 2. Les vecteurs acoustiques sont de dimension $N_c = 39$ et contiennent 12 coefficients MFCC, l'énergie ainsi que les dérivées première et seconde de ces quantités. L'apprentissage de *HMM1* et *HMM2* est obtenu avec 20 itérations de l'algorithme de Baum-Welch. Pour *HMM1*, deux passes de

la procédure itérative ont été utilisées.

L'application des post-traitements par modèle de frontières et par l'algorithme de Brandt a été testée sur les segmentations produites par les deux algorithmes *HMM1* et *HMM2*. Les segmentations produites par post-traitements sur *HMM1* et *HMM2* sont appelées respectivement *Affin1* et *Affin2* tandis que les segmentations obtenues par l'algorithme de Brandt sont nommées *Br1* et *Br2*. Pour *Affin1* et *Affin2*, les paramètres N, M et e présentés sur la figure 1 sont fixés respectivement à 2, 20 ms et 30 ms. Pour *Br1* et *Br2*, l'ordre p des modèles AR est égal à 12 et la longueur minimale des fenêtres w_1 et w_2 est de 10 ms. Pour les algorithmes *HMM2*, *Affin1* et *Affin2* utilisant un apprentissage supervisé, une même partie de la base de données segmentée manuellement et constituée de 100 phrases choisies aléatoirement est utilisée. Afin de pouvoir effectuer une comparaison des différents algorithmes, la base de test considérée est le corpus complet privé de ces 100 phrases. La mesure de qualité choisie est le taux de segmentation correcte pour une tolérance égale à 20 ms, limite jugée acceptable pour garantir une qualité convenable de la voix synthétique.

D'après les résultats du tableau 1, l'algorithme *HMM2* apporte une amélioration significative par rapport à sa version non supervisée HMM1 ; le post-traitement par modèle de frontière semble être le plus performant ; enfin, l'algorithme de Brandt dégrade significativement les résultats des méthodes à base de HMM, qu'il soit appliqué sur les segmentations issues de *HMM1* ou de *HMM2*. Néanmoins, lors de ces expériences, nous avons pu constater que l'algorithme de Brandt parvient à bien localiser certains types de transitions (e.g. parole/silence, non-voisé/voisé). Il nous apparaît donc judicieux de tirer profit des performances de chaque algorithme selon les transitions à traiter. C'est sur la base de cette constatation que nous proposons dans la section suivante une approche par fusion de plusieurs segmentations.

Tab. 1: Taux de segmentation correcte à 20 ms pour chacun des algorithmes

HMM1	*Affin1*	*Br1*
88.29%	90.70%	84.50%
HMM2	*Affin2*	*Br2*
91.85%	92.70%	83.24%

3. Combinaison de plusieurs segmentations

3.1. Principe

Dans cette section, nous proposons un mécanisme permettant de combiner les marques de segmentation produites par différents algorithmes. Le principe de la méthode est d'analyser les performances de chacun des K algorithmes candidats sur différentes classes de transitions phonémiques, de manière à favoriser, lors de l'étape de segmentation, certains algorithmes par rapport à d'autres en fonction des transitions phonémiques à traiter. Plus précisément, étant donné un ensemble $\{c_1, \ldots, c_T\}$ de T classes, il s'agit d'estimer les taux de segmentation correcte $\alpha_k(c_i, c_j)$ à 20 ms pour $(i, j) \in \{1, \ldots, T\}^2$ et $k \in \{1, \ldots, K\}$. Pour mener cette étape d'apprentissage, il est bien entendu nécessaire de disposer de la segmentation manuelle d'une petite partie du corpus.

Ces taux de segmentation correcte vont permettre de combiner ces algorithmes en fonction des transitions à traiter comme suit. Soit s une transition dont les contextes phonétiques gauche et droit sont respectivement $c_g(s)$ et $c_d(s)$. Notons $t_k(s)$ la marque de segmentation de la transition s obtenue par le $k^{\text{ième}}$ algorithme. Une première solution consiste tout simplement à choisir, pour une transition donnée, l'algorithme fournissant en moyenne la meilleure précision, ce qui revient à effectuer avec la terminologie de [7] une fusion dure de la forme :

$$\hat{t}_{dure}(s) = \frac{\sum_{k \in A} t_k(s)}{Card(A)} \quad (1)$$

où A est l'ensemble des algorithmes k qui maximise la quantité $\alpha_k(c_g(s), c_d(s))$ avec $k \in \{1, \ldots, K\}$. Précisons que A ne se résume pas à un seul élément. En effet, si les transitions entre les classes i et j ne sont pas observées dans le corpus d'apprentissage, alors les taux $\alpha_k(i, j)$ ne sont pas définis. Dans ce cas, nous posons $\alpha_k(i, j) = 1$ pour tout k ; on a alors $Card(A) = K$ et l'équation (1) devient une simple moyenne des K marques de segmentation produites par les K algorithmes.

Une autre façon de procéder est d'opérer une fusion douce [7] entre les marques de segmentation issues de chacun des K algorithmes. Cela revient à déterminer l'instant de segmentation comme étant le barycentre suivant :

$$\hat{t}_{douce}(s) = \frac{\sum_{k=1}^{K} \alpha_k(c_g(s), c_d(s)) t_k(s)}{\sum_{k=1}^{K} \alpha_k(c_g(s), c_d(s))}.$$

Notons que si les poids $\alpha_k(i, j)$ sont égaux pour tout k, alors les fusions dure et douce sont équivalentes.

3.2. Expériences et résultats

Dans cette section, nous présentons les résultats obtenus en appliquant les deux stratégies de fusion présentées précédemment d'une part sur le triplet S_1 = (*HMM1*, *Affin1*, *Br1*) et d'autre part sur S_2 = (*HMM2*, *Affin2*, *Br2*). La fusion a été effectuée en considérant les 12 classes suivantes : plosives voisées, plosives sourdes, fricatives voisées, fricatives sourdes, voyelles orales, voyelles nasales, diphtongues, consonnes nasales, consonnes liquides, semi-voyelles, pauses et silences.

Tab. 3: Taux de segmentation correcte à 20 ms pour les différentes stratégies de fusion testées

	Fusion dure	Isobary-centre	Fusion douce	Fusion optimale
$S1$	93.10%	92.80%	93.41%	93.68%
$S2$	93.53%	94.11%	94.65%	94.71%

L'estimation des taux de segmentation correcte utiles pour les fusions dure et douce est faite sur un corpus d'apprentissage constitué de 100 phrases différentes de celles utilisées pour l'apprentissage de *HMM2*, *Affin2* et *Affin1*. Les méthodes de fusion sont ensuite

Tab. 2: Pouvoir de correction des 3 stratégies de fusion

Position des marques en cas d'erreur d'au moins un algorithme	Fréquence d'occurrence	Taux de correction après fusion dure	Taux de correction après fusion isobarycentre	Taux de correction après fusion douce
3 marques du même côté	20.35%	51.25%	48.50%	50.57%
2 marques du même côté	8.11%	79.14%	95.07%	95.71%

évaluées en calculant les taux de segmentation correcte à 20 ms sur le corpus de 7300 phrases privé des phrases utilisées d'une part pour l'apprentissage supervisé des algorithmes et d'autre part pour la détermination des fonctions de fusion douce et dure. Nous comparons également les algorithmes de fusion proposés à deux autres méthodes de fusion : la première dénomée isobarycentre consiste à faire une simple moyenne des instants de segmentation fournis par chacun des 3 algorithmes ; la seconde est une fusion douce optimale en ce sens que les taux de segmentation correctes $\alpha_k(c_i, c_j)$ ont été estimés sur l'ensemble du corpus.

Les résultats consignés dans le tableau 3 montrent tout d'abord que, quelle que soit la stratégie de fusion employée, le taux de segmentation correcte après fusion est toujours supérieur à celui fourni par le meilleur des algorithmes impliqués dans la fusion (*Affin1* et *Affin2*). Par exemple, le taux de segmentation correcte passe de 92.70% pour *Affin2* à 94.65% dans le cas d'une fusion douce entre *HMM2* , *Affin2* et *Br2*, soit une réduction du taux d'erreur de 27%. En outre, la fusion douce se révèle être globalement la plus performante. Elle permet notamment d'améliorer significativement les taux obtenus par la méthode de l'isobarycentre, ce qui valide ainsi l'intérêt de l'apprentissage des $\alpha_k(c_i, c_j)$ opéré. Notons enfin qu'un apprentissage réalisé sur l'ensemble du corpus ne conduit pas à une augmentation très sensible des taux de segmentation correcte.

Le tableau 2 permet d'analyser plus finement le comportement des algorithmes de fusion dans deux configurations. La première correspond au cas où au moins un des algorithmes produit une erreur supérieure à 20 ms et que les trois marques de segmentations estimées sont situées du même côté par rapport à la marque de segmentation manuelle. Dans une telle configuration qui concerne 20.35% des transitions, la marque optimale correspond à celle fournie par l'algorithme ayant produit l'erreur la plus faible et par conséquent une stratégie de fusion dure serait *a priori* plus adaptée. Cependant, les pouvoirs de correction des méthodes de fusion dure et douce sont équivalents et légèrement supérieurs à la fusion par l'isobarycentre. Ceci illustre que dans un tel cas il est difficile de déterminer de manière fiable l'algorithme de segmentation le plus adapté. En revanche, lorsqu'une erreur se produit et que les trois marques de segmentation sont situées de part et d'autre de la marque manuelle, le pouvoir de correction des différentes stratégies de fusion est nettement amélioré. Dans cette deuxième configuration qui représente 8.11% des transitions, les taux de correction des stratégies de fusion douce et isobarycentre sont respectivement de 95.71% et de 95.07%, ce qui montre l'intérêt de procéder à un calcul barycentrique. La stratégie de fusion dure, bien que moins adaptée dans ce cas, parvient tout de même à résoudre 79.14% des erreurs contre 51.25% pour la première configuration.

4. Conclusion

Dans cet article nous avons étudié les performances relatives de trois méthodes de segmentation : l'une globale basée sur le formalisme des HMM et les deux autres locales visant à détecter une rupture au voisinage d'une marque. Nous avons de plus proposé deux stratégies permettant de fusionner les marques de segmentation issues des différents algorithmes. Ces stratégies ont été évaluées sur un corpus de parole dédié à la synthèse vocale et conduisent à une amélioration très sensible des taux de segmentation correcte à 20 ms. Ces méthodes semblent donc prometteuses et seront validées prochainement sur d'autres corpus et pour d'autres langues. Des travaux futurs seront également menés pour traiter le cas où seule une chaîne phonétique approchée obtenue de façon automatique est disponible.

5. Remerciements

Nous remercions Toufic Chmayssani pour ses contributions dans la mise en œuvre de l'algorithme de Brandt.

Références

[1] *The HTK book for HTK V3.0.* 2001.

[2] Y.J. Kim and A. Conkie. Automatic segmentation combining an hmm-based approach and spectral boundary correction. *ICSLP 2002,Colorado*, September 2002.

[3] L.Wang, Y. Zhao, M. Chu, J. Zhou, and Z. Cao. Refining segmental boundaries for tts database using fine contextual-dependent boundary models. *ICASSP*, vol.I :641–644, 2004.

[4] J. Matousek, D. Tihelka, and J. Psutka. Automatic segmentation for czech concatenative speech synthesis using statistical approach with boundary-specific correction. *Eurospeech*, 2003.

[5] S. Nefti. *Segmentation automatique de la parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance.* PhD thesis, Université de Rennes I, 2004.

[6] R.A. Obrecht. A new statistical approach for the automatic segmentation of continous speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.36 :29–40, January 1988.

[7] S. Pigeon. *Authentification multimodale d'identité.* PhD thesis, l'Université Catholique de Louvain, 1999.

Influence des paramètres psycholinguistiques du cocktail party sur la compréhension d'un signal de parole cible

Claire Grataloup[1], Michel Hoen[1, 2], François Pellegrino[1] & Fanny Meunier[1]

1 Laboratoire Dynamique Du Langage CNRS UMR 5596-Université Lyon2
Institut de Sciences de l'Homme 14 avenue Berthelot 69363 LYON Cedex 07

2 Laboratoire Neurosciences et Systèmes Sensoriels
CNRS UMR 5020, Université Claude Bernard, Lyon, France

claire.grataloup@univ-lyon2.fr

ABSTRACT

This paper presents results from an experiment studying the cognitive ability to understand a speech signal in a babble background noise. We further tested subject's sensitivity to characteristics of target and competitor words. Our results show that words are better reconstructed than pseudowords. Intelligibility of words is not influenced by a change (number of voices, frequency of words) in the background babble noise whereas intelligibility of pseudowords is. Pseudowords perception is easier when words that constitute the background noise are low frequency words and when the number of voices is fewer.

1. INTRODUCTION

La perception du langage parlé est une tâche complexe, menée quotidiennement, et qui implique des fonctions cognitives de haut degré. La plupart du temps, l'écoute du signal de parole est endommagée par des artefacts qui perturbent sa compréhension : la présence de bruit ambiant en est l'exemple le plus courant. Le système cognitif doit donc contourner cette difficulté en reconstruisant les portions de signal mal ou non-perçues. Plusieurs études ont montré que dans de telles circonstances la parole reste, dans une certaine mesure, intelligible [1-2-3]. Il existe donc une capacité cognitive de restauration de la parole dégradée.

Une situation particulièrement complexe à traiter et pourtant fréquente se présente lorsque le signal que nous devons percevoir est « camouflé » à l'intérieur d'un flot de paroles provenant de différents locuteurs. Bien que le message cible soit très dégradé, notre système cognitif reste capable de le restaurer de façon suffisante à ce que nous le comprenions. Ce phénomène, décrit par Cherry en 1981 [4] comme l'effet « cocktail party » a été étudié à plusieurs reprises (Bronkhorst, 2000 pour une revue [5]) pour tenter d'identifier les processus cognitifs qui permettent d'isoler la voix qui nous intéresse parmi un brouhaha sonore composé de plusieurs autres voix. Les résultats montrent que la compréhension du message cible dépend à la fois du masquage informationnel et du masquage énergétique imposés par les voix concurrentes [6]. Le masquage énergétique correspond à un recouvrement spectrotemporel même partiel du son cible et du son concurrent. Le masquage informationnel est dû à un recouvrement des informations colportées par les deux signaux.

Une étude de Hoen, Grataloup, Grimault, Perrin, Perrot, Pellegrino, Meunier et Collet (2006) [7] a récemment étudié la sensibilité des locuteurs aux caractéristiques du mélange de parole concurrent. Des mots isolés étaient présentés dans des cocktails de voix composé de 4, 6 ou 8 voix présentées à l'endroit ou inversées (reversed speech). Les résultats révèlent une meilleure performance globale pour la condition à 6 voix que pour celles à 4 et 8 voix. De plus, pour un cocktail composé de 4 voix les mots cibles sont mieux perçus lorsque la parole est inversée que lorsqu'elle est à l'endroit. Cet effet disparaît pour les cocktails 6 et 8 voix. Le masque énergétique augmentant avec le nombre de voix compris dans le cocktail, les auteurs interprètent ces résultats comme révélant qu'à 4 voix dans la condition à l'endroit le masquage informationnel est en place, les mots du cocktail pouvant être activés -ce qui n'est pas le cas pour la condition inversée- et qu'à 6 et 8 voix il disparaît, ne laissant la place qu'au masque énergétique.

Afin d'approfondir cette hypothèse d'activation lexicale des mots du cocktail, nous avons réalisé une expérience ou étaient testés : pour les cibles, le facteur type d'item (mot/pseudomot) avec pour les mots leur fréquence et leur nombre de voisins phonologiques ; et pour le bruit, le nombre de voix et la fréquence des mots qui le constituent. Cette étude mesurait la reconstruction cognitive de signaux de parole (mots et pseudomots) détériorés par la présence de voix concurrentes (cocktail).

2. EXPÉRIENCE

Le principe de l'expérience est de faire entendre à des sujets normo entendants des signaux de parole cibles présentés à l'intérieur d'un cocktail de voix concurrentes. La tâche consiste à identifier l'item cible prononcé par une voix différente de celles composant le bruit de fond.

2.1. Méthode

Matériel : Items cibles

Nous avons sélectionné à l'aide de la base *Lexique* [8] 120 noms communs de la langue française, monosyllabiques et de vocabulaire courant. Deux critères étaient contrastés : leur fréquence d'occurrence dans la langue (facteur f) et leur nombre de voisins phonologiques (facteur v). Ces deux facteurs ont été croisés de façon à construire quatre catégories de 30 mots cibles chacune : table 1. Par exemple, le mot *mage* a une faible fréquence d'occurrence mais possède beaucoup de voisins phonologiques (exemples : *cage, gage, nage, page, rage, sage*…).

Table 1 : Moyennes et fourchettes des fréquences et des voisins phonologiques utilisées

	fréquence	voisins
-	M=2.54 [1, 4.94]	M=11.68 [3, 17]
+	M=66.39 [50.1, 149.23]	M=26.35 [21, 35]

120 pseudomots monosyllabiques ont également été construits en recombinant les phonèmes des mots cibles. Les 240 items ont été enregistrés (22 kH, mono, 16 bits) par une locutrice de langue maternelle française dans un caisson insonorisé. Le matériel utilisé pour l'enregistrement se composait : du logiciel Wavelab lite version 2.53 Steinberg editor, d'une carte son digigram VX pocket440, d'un préamplificateur Behringer ultragain MIC 2000 et d'un micro Rode NT1 équipé d'une membrane Popkiller K&M. Les enregistrements ont ensuite été normalisés à -3 dB à l'aide du logiciel Adobe® Audition® 1.0.

Matériel : Cocktails

Nous avons créé 6 types de cocktails à partir de 16 enregistrements (de 12 min en moyenne) réalisés par 8 locuteurs différents (4 hommes et 4 femmes) qui lisaient chacun une liste de mots fréquents et une liste de mots peu fréquents (1250 mots par liste). Dans les deux listes contrastées en fréquence nous avons équilibré le nombre de lettres des mots, le nombre de syllabes (voir table2) et la proportion de mots de 1, 2, 3 et 4 syllabes.

Table 2 : Critères d'équilibration des listes de mots composant le fond sonore. M = moyenne, ET = écart-type.

Critère	Liste F+	Liste F-
Fréquence	M=151.25 ET=451	M= 0.45 ET=0.3
Nb lettres	M=7.45 ET=2	M=7.81 ET=1
Nb syllabes	M=2.45 ET=1	M=2.63 ET=1

Nous avons ainsi constitué des cocktails C_4 composés de 2 voix féminines et de 2 voix masculines, des cocktails C_6 composés de 3 voix féminines et de 3 voix masculines et des cocktails C_8 composés de 4 voix féminines et de 4 voix masculines. Chaque cocktail existe en 2 versions, l'une fréquente (F+) et l'autre peu fréquente (F-). Dans chaque cocktail, nous avons découpé 120 extraits d'une durée de 4 secondes chacun à l'aide du logiciel Matlab qui nous a permis également de générer les stimuli finaux en superposant chaque item cible avec un extrait de chaque cocktail.

Listes expérimentales

Chaque item cible a été superposé à un extrait de chacun des 6 cocktails existants. Au total, nous avons donc généré 240*6=1440 stimuli. Nous avons ensuite créé 6 listes de mots et 6 listes de pseudomots comportant chacune 120 items. Les 6 versions de chaque item ont été réparties dans les 6 listes et contrebalancées de façon à ce que chaque item n'apparaisse qu'une seule fois par liste.

Procédure expérimentale

Les participants étaient placés face à un écran d'ordinateur de type PC, ils portaient un casque audio (Beyerdynamic DT 48) qui diffusait les stimuli un à un en mode binaural. Une consigne spécifique soit aux mots soit aux pseudomots leur était donnée oralement puis réapparaissait à l'écran en début d'expérience. Chaque sujet a été confronté à l'une des 6 listes de mots et à l'une des 6 listes de pseudomots dont l'enchaînement était spécifique à chaque sujet. La phase de test était précédée par une phase d'entraînement.

Chaque stimulus se compose de 4 secondes de cocktail à l'intérieur duquel l'item cible apparaît 2.5 secondes après le début du bruit. Après chaque stimulus ils devaient retranscrire au clavier l'item cible perçu. La moitié des sujets a commencé par les mots et l'autre moitié par les pseudomots. Une pause était effectuée entre les deux moitiés de l'expérience qui durait 45 minutes.

Sujets

Quarante sujets (25 femmes, 15 hommes) de langue maternelle française ont passé l'expérience1. Leur âge variait entre 18 et 25 ans (moyenne=21.5 ans). Tous ont passé une audiométrie tonale confirmant une audition normale (seuils<20dB) sur la gamme de fréquence des sons de la parole humaine. Aucun d'entre eux ne souffrait de troubles du langage et tous avaient une vue normale ou corrigée. Les participants étaient naïfs quand au but de l'étude et ont été rémunérés 7.5 € chacun.

2.2. Résultats

Nous avons effectué une analyse statistique ANOVA sur les 40 sujets et 240 items en considérant comme variable aléatoire d'une part, les sujets (*F*1) et d'autre part les items (*F*2). La variable dépendante était le % de restitution entière et correcte des items par les sujets.

Effet des items cibles

On observe d'une manière très forte que les mots sont mieux restitués 61% (ET=5.56) que les pseudomots

39.42% (ET=7.02) : figure1. $F1(1,39)=294.87$; $p<.0001$; $F2(1,238)=34,73$; $p<.0001$.

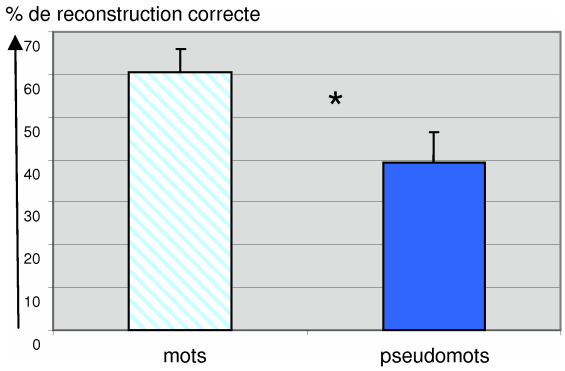


Figure 1 : Effet du type d'item cible sur le % de reconstruction. L'étoile signifie que la différence est statistiquement différente au seuil .05.

Pour les mots on observe un effet simple de leur fréquence d'occurrence. Les mots de haute fréquence sont mieux reconstruits 71% (ET=6.39) que les mots de basse fréquence 50% (ET=6.73). $F1(1,39)=368.72$; $p<.001$; $F2(1,118)=14.92$; $p<.001$. On observe également un effet simple du nombre de voisins phonologiques des mots cibles. Contrairement à ce à quoi on se serait attendu, les mots qui ont le plus de voisins sont ceux qui sont le mieux reconstruits 70% (ET=6.30) pour la condition v+ contre 51% (ET=6.95) pour la condition v- : figure2. $F1(1,39)=249.15$; $p<.0001$; $F2(1,118)=14.46$; $p<.001$.

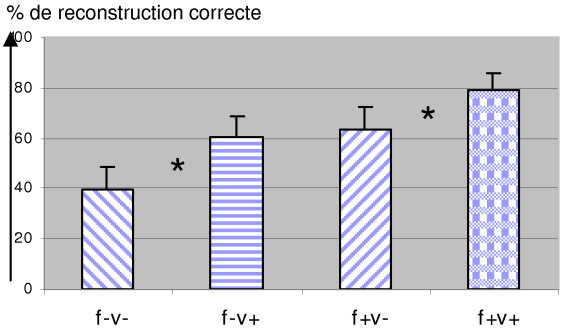


Figure 2 : Effet de la fréquence et du nombre de voisins phonologiques des mots cibles sur leur % de reconstruction.

Effet des cocktails

On n'observe pas d'effet significatif du nombre de voix des cocktails. En revanche on observe un effet de la fréquence des mots du cocktail sur la restitution des pseudomots uniquement : figure 3. En moyenne, les pseudomots sont reconstruits à 38% dans un cocktail fréquent (ET=7.85) et à 41% dans un cocktail non fréquent (ET=9.21). $F1(1,39)=3.98$; $p=.05$, $F2(1,119)=4.75$; $p<.05$. Les mots sont respectivement reconstruits à 61% (ET=7.44) et 60%(ET=7.37). $F1(1,39)=1.4$; n.s. ; $F2(1,119)=2.7$;n.s..

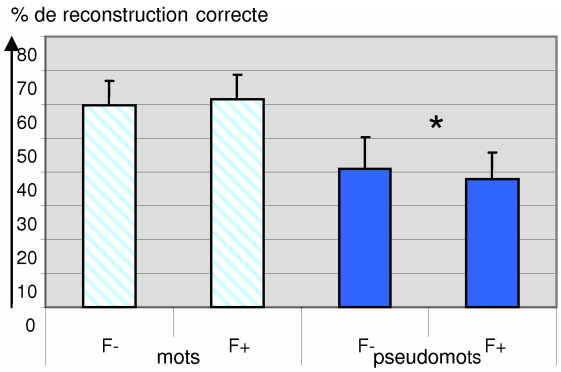


Figure 3 : Effet de la fréquence des mots du cocktail sur le % de reconstruction des items cibles.

Interaction nombre de voix et fréquence des mots du cocktail

Il faut noter que nous observons une interaction significative entre le nombre de voix et la fréquence des mots du cocktail pour la reconstruction des items. Pour les mots, cette interaction n'est significative que par items : $F1(2,78)=2.1$;n.s. ; $F2(2,238)=3.37$, $p<.05$. Le traitement des pseudomots en revanche présente une interaction significative par sujets $F1(2,78)=3.7$, $p<.05$ et par items $F2(2,238)=4.64$; $p=.01$. En moyenne, on n'observe pas de différence entre cocktails pour la condition 6 voix (42% (ET=14,1) contre 38% (ET=10.43)) ni pour la condition 8 voix (36%, ET=12.8 contre 38%, ET=12.07). Cependant, on observe une différence significative entre les % de reconstruction pour les deux types de cocktails dans la condition à 4 voix. Les pseudomots sont reconstruits à 44% (ET=13.6) dans le cocktail peu fréquent et seulement à 38% (ET=13.77) dans le cocktail fréquent : figure 4.

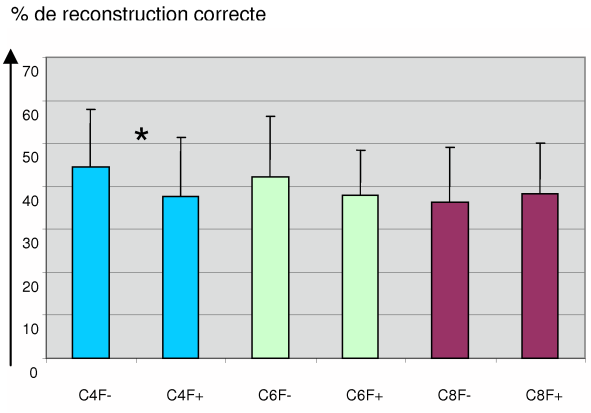


Figure 4 : Effet d'interaction entre la fréquence et le nombre de voix du cocktail sur le % de reconstruction des pseudomots cibles.

3. DISCUSSION

Les résultats montrent que les mots sont mieux reconstruits que les pseudomots ce qui est cohérent car les pseudomots n'ont pas de représentation lexicale stockée en mémoire. De ce fait, ils ne bénéficient d'aucune aide lexicale au moment de l'effort de reconstruction contrairement aux mots qui, eux, bénéficient de cette aide.

En ce qui concerne les mots, on observe un fort effet de leur fréquence d'occurrence. Les mots les plus fréquents sont mieux restitués que les mots de basse fréquence. La fréquence du mot influence les performances de restitution quelque soit le nombre de voix qui composent le bruit de fond et quelque soit la fréquence des mots du bruit de fond. L'effet de fréquence est un effet très robuste qui apparaît dans la plupart des tâches cognitives proposées aux sujets [9]. La fréquence étant une caractéristique de stockage du mot dans le lexique mental, plus sa fréquence est élevée, plus l'accès au mot est facile. On observe ici, que cet effet se retrouve lorsque la perception des mots cibles est perturbée par celle de mots concurrents.

Le résultat le plus intéressant observé dans cette expérience est sans aucun doute l'effet de fréquence du cocktail sur la restitution des mots cibles. Lorsque les mots distracteurs du cocktail sont de basse fréquence, le % de reconstruction est plus élevé. A l'inverse, si les mots du cocktail sont de forte fréquence, ils gênent la reconstruction du mot cible. Ce résultat peut-être interprété de deux façons : soit par un effet attentionnel différent selon le niveau de fréquence des mots, soit par une différence d'activation des mots des deux catégories. En d'autres termes : lorsque les mots du cocktail sont de forte fréquence, ils attirent plus l'attention du locuteur (effet de familiarité) et de ce fait les ressources attentionnelles disponibles pour traiter l'item cible sont moindres (diminution du % de reconstruction) ou bien, lorsque les mots du cocktail sont de basse fréquence, ils sont moins saillants et le système dispose donc de plus de ressources attentionnelles pour traiter le stimulus cible. L'autre explication, peut-être plus plausible, est que les mots de basse fréquence du cocktail sont moins activés et donc moins en compétition avec les mots cibles. D'autres expériences sont nécessaires afin de clarifier ce point. Cependant, il est certain que les locuteurs sont sensibles aux caractéristiques lexicales des mots du cocktail.

De plus, bien que l'on n'observe pas d'effet simple du nombre de voies du cocktail, on observe cependant une interaction entre la fréquence des mots et le nombre de voix composant le cocktail. L'effet de fréquence des mots du cocktail n'apparaît que pour la condition 4 voix : Les stimuli sont significativement mieux reconstruits dans un cocktail peu fréquent à 4 voix que dans un cocktail fréquent à 4 voix. Ce résultat suggère que c'est bien dans la condition où seulement 4 voix sont mélangées que les locuteurs peuvent être sensibles à la qualité des mots prononcés. Au-delà de 4 voix, le bruit de fond devient trop dense pour pouvoir discerner une différence de fréquence entre les mots composant les deux types de cocktails. A 4 voix cependant, le bruit de fond n'est pas encore suffisamment chargé et il est possible que les locuteurs soient influencés par un facteur lexical des mots concurrents. Ce résultat rejoint celui de Hoen et collaborateurs [7].

4. CONCLUSION

Cette étude présente les premiers résultats mettant en évidence une sensibilité des locuteurs aux caractéristiques lexicales de voix concurrentes de type cocktail lors d'une tâche de perception de parole. Ce paradigme pourrait permettre alors l'exploration des compétitions lexicales entrant en jeu dans la compréhension de la parole d'une manière plus simultanée que les paradigmes d'amorçages actuellement utilisés.

5. NOTES DES AUTEURS ET REMERCIEMENTS

Nous remercions la région Rhône-Alpes qui a permis la réalisation de cette étude grâce au projet Emergence 2004 attribué à Fanny Meunier.

BIBLIOGRAPHIE

[1] Warren, R.M. (1970). Restoration of missing speech sounds. Science, 167, 392--393.

[2] Scott, S. K., Blank, S. C., Rosen S., and Wise, R. J. S. (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400-06.

[3] Davis, M.H. & Johnsrude, I.S. (2003). Hierarchical processing of spoken language comprehension. Journal of Neuroscience 23, 3423-3431.

[4] Cherry, E. (1953). ''Some experiments on the recognition of speech, with one and two ears,'' J. Acoust. Soc. Am., 25, 975–979.

[5] Bronkhorst, A. (2000). ''The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,'' Acustica 86, 117–128.Trans. Speech and Audio Proc., 7(6):697-708, 1999.

[6] Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott K.R. (2001). ''Informational and energetic masking effects in the perception of multiple simultaneous talkers,'' J. Acoust. Soc. Am., 110, 2527–2538.

[7] Hoen, M.; Grataloup, C.; Grimault, N.; Perrin, F.; X. Perrot; Pellegrino, F.; Meunier, F.; Collet, L. Tomber le masque de l'information : effet *cocktail party*, masquage informationnel et interactions psycholinguistiques en situation de compréhension de la parole dans la parole. JEP 2006.

[8] New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE. L'Année Psychologique, 101, 447-462.

[9] Segui, J., Melher, J., Frauenfelder, U., et Morton, J. (1982). The word frequency effect and lexical access. Neuropsychologia, 20(6), 615-627

Analyse des stratégies de chunking en interprétation simultanée

Myriam Piccaluga & Bernard Harmegnies

Université de Mons-Hainaut
Institut de linguistique, 22 place du Parc, 7000 Mons, Belgique
il@umh.ac.be
http://www.umh.ac.be/linguistique.html

ABSTRACT

In this paper, which is meant as a methodological account, we focus on a new variable ("Ecart Inter Syllabique": EIS), intended to improve the study of speech chunks produced by subjects performing a task of simultaneous interpreting ("IS"). The variable is introduced on the basis of a discussion of the main methodological trends in the field, with the aim of improving the validity and reliability of the numerical treatments applied to the study of IS. An experimental essay is performed on a prototypical sample of 4 subjects, performing IS under several conditions. The behaviour of the variable within the design suggests its interest for future research.

1. INTRODUCTION

L'interprétation simultanée est une tâche consistant, pour le sujet, à réémettre dans une langue un message qu'il est en train de recevoir dans une autre. Cette activité de traduction orale « online » peut, comme toute activité de médiation langagière, être observée en tant que comportement spontané chez tout bilingue ; elle constitue alors un cas particulier de *traduction naturelle* [1]. Eduquée à la faveur d'une formation intensive, elle peut se développer dans l'exercice professionnel de l'*interprétation de conférence* ou de l'*interprétation de liaison*, prestations professionnelles dont l'internationalisation du monde du travail a contribué, durant les dernières décennies, à accroître la visibilité. Notre objet d'étude principal est cette tâche – nous l'appellerons ici *la tâche interprétative :* TI - quel qu'en soit le contexte d'effectuation et quelle que soit l'expertise des sujets l'exécutant.

Les différentes études cognitives, tant empiriques que spéculatives, qui ont abordé la TI [e.g., 2,3] s'accordent quant à l'idée que la *simultanéité* qui caractérise la tâche est plus complexe que la simple juxtaposition temporelle d'un processus de réception et d'un processus d'émission. Pour Paradis [4], la TI suppose l'effectuation successive de nombreuses sous-tâches permettant d'assurer la compréhension, la traduction et le monitoring (saisie en mémoire échoïque d'un segment de discours source, décodage, représentation non linguistique du message, encodage en langue cible, émission en langue cible, stockage de la production en mémoire échoïque, représentation non linguistique du message, comparaison des deux représentations non linguistiques). Un modèle de ce genre postule évidemment la segmentation du signal d'entrée en portions suffisamment réduites pour permettre le traitement. Une chaîne de sous-tâches peut ainsi être mise en route alors qu'une autre est déjà active : plusieurs chaînes de traitement sont ainsi susceptibles de fonctionner en parallèle à condition que, en un temps t déterminé, les différentes sous-tâches actives simultanément soient de nature différente et portent sur des portions différentes du discours source. Cette idée d'un nécessaire découpage du discours d'entrée se retrouve chez d'autres auteurs, qui, avec la notion de *chunk*, réfèrent à une portion du discours source caractérisée par le fait qu'elle constitue une unité brève, mais suffisamment riche d'informations pour faire l'objet d'un traitement sémantique unitaire. Cette vision est à lier avec la présence, dans plusieurs modèles, d'un *buffer* mémoriel, qui évoque d'ailleurs, explicitement ou non, les dispositifs de stockage d'entrée que comportent les modèles plus généraux du langage [5].

Il est raisonnable de penser que l'aptitude du sujet à effectuer un chunking efficace du discours source est une condition *sine qua non* (mais non suffisante) de l'obtention d'un produit interprétatif de qualité. L'étude des capacités du sujet à ce type de découpage apparaît donc fort intéressante en vue de l'analyse des facteurs influant sur l'efficacité du sujet dans la tâche, partant, des processus cognitifs sollicités et, in fine, de la validité des modèles de la TI.

L'étude des stratégies de chunking n'est cependant pas aisée. En effet, le chunking n'est à l'origine d'aucun comportement directement observable. Une alternative intéressante à des designs complexes risquant de dénaturer la tâche peut cependant consister en la sollicitation du sujet dans le cadre d'une TI simple, avec, dans le chef du chercheur, une focalisation non pas sur l'*entrée* de la *boîte noire* mais bien sur la *sortie*. Si le traitement se déroule sans heurts, on peut en effet s'attendre à ce que le chunking d'entrée soit régulier et à ce qu'il induise un chunking de sortie lui aussi régulier. Par contre, un chunking d'entrée irrégulier ou toute perturbation, à quelque niveau de traitement que ce soit, rend probable un chunking de sortie irrégulier. L'idée qui se dégage de ces réflexions amène donc à

prendre en compte l'analyse du chunking de sortie en tant que révélateur indirect du chunking d'entrée.

Dans cet article, nous proposons une contribution à caractère exploratoire et méthodologique visant à étudier l'intérêt d'une variable (l'Ecart Inter Syllabique : EIS), escomptée apte à rendre compte des phénomènes de chunking sans nécessiter le recours à des procédures classiques dont nous montrons les limitations. A cet effet, la variable à l'étude est testée dans le cadre d'un essai expérimental exploratoire visant à en investiguer les qualités métrologiques.

2. LA QUANTIFICATION DU PHENOMENE

2.1. Des approches classiques

L'étude de la structuration temporelle du discours cible peut, en première analyse, se baser sur l'une ou l'autre de deux voies alternatives. La première se centre sur l'évaluation de l'importance relative du signal articulé par rapport au temps de la locution, soit sur base de dénombrements, soit au départ de mesures de durée des groupes syllabiques, dans la foulée des travaux fondateurs de Goldman-Eisler [e.g. 6]. La deuxième recourt à l'analyse des pauses, celles-ci pouvant être considérées comme délimiteurs de chunks.

Que l'on examine la locution avec un regard centré sur les sections de signal de parole qu'elle comporte ou qu'on la considère avec une attention plus ciblée sur ses lacunes, voire qu'on adopte simultanément les deux perspectives, on bute inévitablement sur un problème délicat, celui de l'établissement des frontières de segments.

Ainsi, il n'est ici pas toujours aisé de délimiter avec certitude l'endroit où finit une syllabe et où commence la suivante. Tel est le cas si, par exemple, la première syllabe se termine par une voyelle et la suivante débute par une voyelle ; tel est également le cas pour certaines structures consonantiques (en particulier les approximantes), qui posent également des problèmes de démarcation.

Quant à l'étude des pauses, si elle peut paraître séduisante, force est de reconnaître qu'elle se heurte cependant à plusieurs difficultés méthodologiques. En effet, l'étude des chunks basée sur l'analyse des pauses considérées comme marqueurs de frontières nécessite la détermination d'un seuil de durée : seules doivent être prises en considération les pauses significatives d'une activité cognitive de plus haut niveau que l'implémentation phonétique. Ceci nécessite que, dans l'univers des pauses, on établisse deux sous-ensembles. Le critère le plus couramment utilisé est celui de la durée qui, pour d'aucuns, doit être supérieure à 250ms [e.g. 6]. Cette valeur est cependant loin d'être la seule que propose la littérature, ce qui est probablement dû, en partie, à des différences méthodologiques entre auteurs, mais peut-être, plus profondément, à la nature-même des pauses, dont les travaux les plus récents montrent bien la diversité [7,8]. De ce point de vue, le modèle taxinomique binaire sous-jacent à la prise de décision dichotomique sur base d'une seule valeur démarcative pose autant problème sur le plan épistémologique que sur le plan méthodologique. Enfin, il faut bien voir que l'analyse d'un phénomène (les chunks) sur base d'un non-événement (la rupture du flux phonatoire) est, par-delà ses limitations épistémologiques, à l'origine de problèmes techniques. Ainsi, si l'on tente d'automatiser un processus de délimitation de chunks, on se heurte rapidement à un autre problème de seuil concernant, cette fois-ci, l'intensité. Si les signaux ont été recueillis dans un environnement un tant soit peu écologique, les phases silencieuses risquent fort d'être contaminées par les bruits environnementaux, ce qui constitue une nouvelle source d'invalidation de la procédure.

Ces différentes approches, qui ont montré leurs limites, se caractérisent de surcroît par le fait qu'elles ne sont que difficilement automatisables, vu la quantité importante de décisions et de spéculations que requiert leur mise en œuvre.

2.2. Une approche innovante

La variable dont nous souhaitons ici tester la qualité (l'EIS) est obtenue par différenciation des moments où apparaissent les pics d'intensité des noyaux syllabiques [9]. Pour chaque pic, on calcule l'intervalle de temps le séparant du sommet syllabique précédent. Pour tout corpus de N syllabes, on obtient ainsi une liste chronologiquement ordonnée de N-1 durées inter syllabiques.

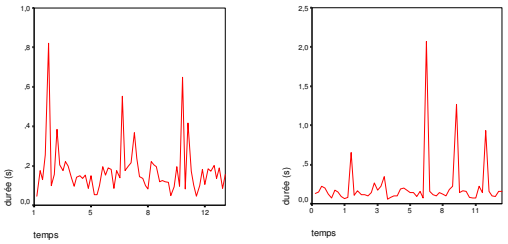


Figure 1 : évolution des valeurs d'EIS (en ordonnée) en fonction du temps (en abscisse), durant les 13 premières secondes de l'effectuation d'une TI par le même sujet, sous conditions peu perturbée (à gauche) et fortement perturbée (à droite).

Comme le montre la figure 1, les graphes d'évolution de l'EIS montrent une alternance plus ou moins régulière de pics et de vallées. Les pics correspondent aux EIS longs, signalent une baisse momentanée du débit ; les vallées, au contraire, correspondent aux périodes caractérisées par un débit plus rapide. C'est donc une forme de graphe d'évolution du débit instantané au cours du temps que nous avons ici. On peut reconnaître, dans l'alternance entre d'une part des zones assez longues caractérisées par des durées

localement faibles et peu variables, et d'autre part, des zones plus brèves caractérisées par un débit local ralenti, le séquencement du flux de parole en chunks de matériau phonique.

Lorsque le processus d'interprétation se déroule sans heurts (partie gauche de la figure 1), l'évolution de l'EIS prend la forme d'une alternance plutôt régulière de pics (les frontières de chunks) et de vallées. Par contre, lorsque le rapport *expertise/difficulté de la tâche* décroît, la régularité est perturbée : on voit apparaître des zones à EIS stablement bas (débit élevé : le sujet tente de rattraper le temps perdu) et d'autres à pics élevés et fréquents (accumulation de pauses longues : le processus interprétatif trébuche).

3. ESSAI EXPERIMENTAL

3.1. Sujets

Les sujets sont au nombre de 4, tous originaires de la région de Barcelone et y résidant depuis plusieurs décennies. Tous ont un haut niveau de maîtrise des langues française et espagnole, qu'ils exercent toutes deux fréquemment; pour chacun, l'espagnol fait cependant figure de langue dominante. Hormis ces similarités, les sujets se caractérisent par des différences en termes de leurs expertises d'une part au plan de la maîtrise de la TI et d'autre part au plan de la maîtrise linguistique (Cfr table 1).

Table 1 : ventilation des sujets en termes d'expertise linguistique (bilingues *vs* quasi-natifs) et d'expertise de la tâche (professionnels *vs* non professionnels).

	Professionnels	Non professionnels
Bilingues	*Int1*	*Étud*
Quasi natifs	*Int2*	*Biling*

3.2. Corpus sources et combinaisons linguistiques

Chaque sujet a été soumis à 6 tâches interprétatives, chacune consistant en le traitement d'une conférence originellement prononcée à la tribune du Parlement Européen. Chaque sujet a dû interpréter 3 discours dans chacune des deux combinaisons (*français vers espagnol* et *espagnol vers français*).

3.3. Perturbations apportées aux corpus sources

Chaque discours source a subi des perturbations introduites en laboratoire. Celles-ci consistaient d'une part en une altération locale du débit (accroissement du débit par réduction à 80%, 70% ou 60% de la durée totale de production sans modification des caractéristiques de F_0) et d'autre part en l'adjonction locale d'un bruit blanc parasitant (0 dB, 3 dB ou 6 dB re/niveau moyen du signal du discours source). Les analyses présentées ici ne portent que sur certaines des portions de discours, caractérisées, chacune, à la fois par un traitement de parasitage et un traitement de compression temporelle.

3.4. Méthodologie statistique

Nous analysons ici les variations de l'EIS au moyen de procédures d'analyse de variance permettant de tester l'hypothèse d'un effet de nos variables indépendantes (désormais « VI ») compte tenu de leurs interactions (dont nous ne pouvons ici rendre compte, par manque de place). Nous nous interrogerons donc sur l'action des VI *parasitage*, *compression temporelle*, *combinaison linguistique* ainsi que *sujet*.

3.5. Résultats

La procédure décrite ci-dessus a abouti au recueil de 12930 valeurs, présentant une moyenne de 268ms, supérieure à la médiane (173ms), ce qui trahit une dissymétrie droite prononcée, (coefficient de 7,560), allant de pair avec une forme fortement leptocurtique (coefficient d'aplatissement de 96,462). Les percentiles 50, 90 et 95 sont d'ailleurs tous trois contenus dans l'intervalle [0-1], alors que l'étendue de la distribution est de près de 10 unités. Etant donné cet évident écart par rapport au modèle gaussien, nous avons appliqué, en vue de l'analyse de variance, une transformation argument-tangente hyperbolique visant à normaliser la distribution. Cette opération s'est révélée globalement satisfaisante, puisqu'elle a permis une diminution du coefficient de dissymétrie de 7,560 à 1,219 et du coefficient d'aplatissement de 96,462 à 3,091 (Cfr fig. 2)

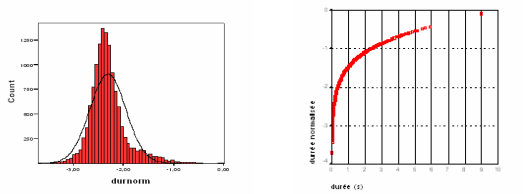


Figure 2 : distribution des durées d'EIS après normalisation ; à gauche : valeurs observées et courbe normale ajustée ; à droite, relation entre les valeurs normalisées (en ordonnée) et les valeurs d'origine.

Afin d'en faciliter la lecture, nous présentons dans les graphiques suivants non les valeurs d'EIS transformées, mais bien les valeurs dans l'unité originale (ms).

Comme le montre la figure 3, les sujets se différencient fortement en termes d'EIS *($F=70,104$, d.l.=3, $\alpha<.001$)*. On voit apparaître deux groupes distincts : d'une part, *biling* et *int2*, qui se caractérisent par des valeurs comprises entre 225 et 250ms et, d'autre part, *étud* et *int1* qui, tous deux, ont des valeurs avoisinant 310ms. Une distinction nette en termes d'expertise de la langue apparaît donc ici. Sous l'effet de la combinaison, les valeurs de l'EIS se modifient aussi. La différence, ténue mais significative *($F=216,368$, d.l.=1, $\alpha<.001$)*,

va dans le sens d'un accroissement lorsque la production est en français, la langue globalement la moins bien maîtrisée dans l'échantillon.

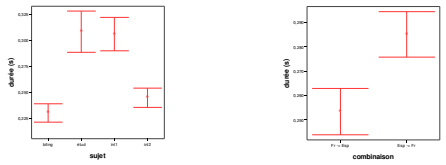


Figure 3 : évolution des valeurs d'EIS (en ordonnée) ; à gauche, en fonction du locuteur (de gauche à droite : *biling, étud, int1, int2*) ; à droite, en fonction de la combinaison (de gauche à droite : *Fr vers Esp*, *Esp vers Fr*).

Comme le montre la figure 4, l'accroissement du taux de parasitage induit une élévation sensible de l'EIS *(F=20,489, d.l.=2, α<.001)*. Celle-ci est quasiment linéaire. Tout aussi linéaire est, au contraire, la décroissance significative *(F=19,474, d.l.=2, α<.001)* que produit l'accroissement du taux de compression, à l'opposé de ce que l'on constate concernant le parasitage.

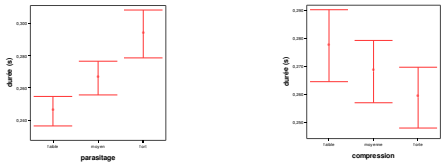


Figure 4 : évolution des valeurs d'EIS (en ordonnée) ; à gauche, en fonction du parasitage (de gauche à droite : *croissant*) ; à droite, en fonction de la compression (de gauche à droite : *croissante*)

4. CONCLUSIONS

Au plan strictement méthodologique, on peut constater que la variable à l'étude s'est montrée apte à différencier, à posteriori, des groupes de mesures correspondant à des classes à priori du dispositif expérimental (classes d'expertise linguistique, de combinaison linguistiques, de perturbations). En première analyse, il semble donc que ses intérêts théoriques (indépendance de toute logique de seuils, robustesse face aux environnements acoustiques adverses) et technique (possibilité d'automatisation aisée de la mesure) s'assortisse d'un intérêt patent en matière d'informativité.

Les observations relevées à propos des VI *sujet*, *combinaison* et *parasitage* vont dans le sens d'un EIS bas pour une situation où les contraintes sont minimes : bonne maîtrise de la langue d'émission et de la langue de réception, absence de bruitage. La variable *compression temporelle* tranche évidemment avec les autres ; il faut bien voir, cependant, qu'elle est la seule à exercer un effet direct sur le séquencement de la réception (et, en conséquence, de l'émission, dans le contexte d'une TI), ce qui confirme en quelque sorte à contrario la sensibilité de la variable.

Notons, par ailleurs, que l'exploitation des données nous a conduits à nous centrer sur un seul aspect des mesures : leur *tendance centrale*. D'autres développements, plutôt centrés sur la *variabilité* de l'EIS, pourraient opportunément être envisagés. Par ailleurs, une étude plus approfondie des propriétés métrologiques, poussant notamment plus avant la recherche d'un caractère gaussien de l'EIS, est probablement souhaitable. Enfin, si notre essai expérimental nous a conduits à privilégier un regard macroscopique visant surtout à éprouver les qualités métrologiques d'un dispositif numérique, une étude plus clinique interrogeant microscopiquement les relations entre les variations de l'EIS et les comportements observés serait de nature à mieux éprouver sa validité ; elle pourrait participer d'une analyse investiguant les plans de la production (aspects phonétique-phonologique, lexical, morpho-syntaxique, pragmatique, etc.), de la médiation (aspects traductologiques), voire des qualités subjectives du discours cible du point de vue du récepteur final (aspects liés à l'intelligibilité).

BIBLIOGRAPHIE

[1] A.M.B. De Groot, The cognitive study of translation and interpretation. Three approaches, in J.H. Danks, G.M. Shreve, S.B. Fountain, M.K. McBeath, *Cognitive processes in translation and interpreting,* Sage Publications, Thousand Oaks, 25-56, 1997.

[2] D. Gerver, Empirical studies of simultaneous interpretation: a review and a model, in R.W. Brislin, (Ed.), *Translation: Applications and Research*, Gardner Press, New York, 165-207, 1976.

[3] B. Moser, Simultaneous interpretation: a hypothetical model and its practical application, in D. Gerver and W. Sinaiko (Eds.), *Language interpretation and communication*, Plenum Press, New York and London, 353-368, 1978.

[4] M. Paradis, Toward a neurolinguistic theory of simultaneous translation: the framework, *International Journal of Psycholinguistics*, 10, 3(29), 319-335, 1994.

[5] A. Baddeley, *Human Memory Theory and Practice (Revised Edition)*, Allyn & Bacon, Boston, 1997

[6] F. Goldman-Eisler, Segmentation of input in simultaneous translation, *Journal of Psycholinguistic Research*, vol. 1, n°2, 127-140, 1972.

[7] E. Campione & J. Véronis, Pauses et hésitations en français spontané, *Actes des 25èmes Journées d'Etudes sur la Parole (JEP 2004)*, 109-112, 2004.

[8] M. Candéa, *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse de doctorat, Université Paris III, 2000.

[9] M. Piccaluga, *Approches psycholinguistiques de l'interprétation*, Thèse de doctorat sous la direction du prof. J.-L. Nespoulous, UMH, 2004.

Produit multiéchelle pour la détection des instants d'ouverture et de fermeture de la glotte sur le signal de parole

Aïcha Bouzid et Noureddine Ellouze

Laboratoire signal, image et reconnaissance de formes (LSTS-ENIT)
Le Belvédère, B. P. 37, 1002, Tunis
bouzidacha@yahoo.fr, N.Ellouze@enit.rnu.tn,

ABSTRACT

This paper deals with robust singularity detection in speech signal using multiscale product method. These singularities correspond to opening and closure instants of the glottis (GOIs and GCIs). Multiscale product method consists of computing the products of wavelet transform coefficients of the speech signal at appropriate adjacent scales. As wavelet modulus maxima are a tool for signal edge detection, first derivative of a Gaussian function, is used for detecting speech signal discontinuities. Speech Multiscale products enhance edge detection. The proposed method is evaluated comparing to the EGG signal references using the Keele University database. This method gives excellent results concerning GOI and GCI detection from speech signal.

1. INTRODUCTION

Les vibrations quasipériodiques des cordes vocales forment l'essentiel de l'excitation du conduit vocal en parole voisée. Parmi les événements qui caractérisent le cycle du larynx, les instants d'ouverture et de fermeture de la glotte occupent une place importante. Ces instants sont notés GCI pour la fermeture (Glottal Closure Instant) et GOI pour l'ouverture (Glottal Opening Instant). L'excitation majeure du conduit vocal se produit à l'instant de fermeture de la glotte. Les cordes vocales se ferment brusquement et restent fermées pour à peu près la moitié du cycle (phase fermée), l'écoulement d'air s'arrête rapidement, ceci entraîne une discontinuité de l'onde du débit glottique appelée aussi flux glottal. Le GCI représente l'instant le plus significatif lorsqu'un point de référence périodique est recherché, car c'est l'événement le plus simple à détecter dans une période du pitch. Tout aussi important est le GOI qui sépare la phase ouverte de la phase fermée mais de façon plus lente et régulière, c'est l'événement le plus difficile à repérer. La connaissance des instants d'ouverture et de fermeture permet de déterminer divers paramètres relatifs à la qualité du son dont principalement la période du pitch T_0 et le quotient ouvert O_q durée de la phase ouverte relative à la période du signal. Le traitement de parole utilise les GCIs et les GOIs ; par exemple, lors de la phase fermée de la glotte, pour déterminer avec précision les paramètres du conduit vocal, tels que les formants [1]. De même l'analyse pitch synchrone nécessite naturellement une connaissance robuste des instants d'ouverture et de fermeture de la glotte. Dans les applications de synthèse vocale, les performances des méthodes PSOLA [2] dépendent fortement de la précision de localisation des marqueurs de pitch. Enfin il est également possible de disposer d'une meilleure caractérisation de la source par une analyse du signal qui se réfère aux GCIs et GOIs. Le signal d'excitation du conduit vocal est porteur d'informations caractéristiques sur la source vocale, l'analyse de ce signal permet de repérer aisément les instants d'ouverture et de fermeture de la glotte. Cette simplicité devient une tâche ardue si l'on cherche à détecter ces instants sur le signal de parole, en effet l'interaction de l'excitation glottique et les variations temporelles de la forme du conduit vocal rendent le signal complexe. Il est ainsi difficile d'isoler l'excitation majeure dans une période du pitch sur le signal de parole. Généralement, dans une période du pitch, on peut noter la présence de plusieurs excitations dont certaines sont significatives. Même si l'instant de fermeture de la glotte correspond à l'excitation majeure dans le cas d'une voyelle normale, en faible voisement, cet instant devient difficile à définir. A l'instant d'ouverture, l'excitation est de nature plus douce donc plus difficile à repérer et pratiquement impossible à détecter sur le signal de parole [1]. Nombreuses, sont les recherches qui tentent d'identifier avec la meilleure précision possible, les instants de fermeture de la glotte GCIs sur le signal d'excitation et le signal de parole, afin de disposer d'une référence périodique et d'une estimation de la période instantanée du pitch. Toutefois, les travaux sur l'instant d'ouverture sont relativement modestes. Ceci s'explique d'une part par la position d'importance première accordée à l'instant de fermeture et d'autre part aux difficultés associées à la détermination de l'instant d'ouverture. Les méthodes proposées pour la détection des instants de fermeture de la glotte à partir du signal de parole sont essentiellement basées soit sur une modélisation LPC de l'erreur comme le résidu ou le filtrage de Kalman [3]; soit sur les discontinuités spectrales observées sur les représentations temps fréquence ou temps échelle. La détection des instants d'ouverture est plus aisément obtenue sur le signal électroglottogramme (EGG) [4], cette approche n'est pas commode dans les applications [5].

Durant ces dernières décennies, la transformée en ondelettes a été fréquemment utilisée pour la détection des singularités d'un signal [6]. La transformée en ondelettes a été aussi utilisée par différents algorithmes pour la détection du pitch [7], [8]. De même l'analyse des coefficients en ondelettes a permis la détermination des instants de fermeture de la glotte [9]. En effet, les instants de fermeture de la glotte correspondent la plupart du temps, à des variations brusques ou des singularités dans le signal de parole.

Ce papier présente une approche simple et robuste de détection des instants d'ouverture et de fermeture de la glotte basée sur le produit des coefficients de la transformée en ondelettes continue du signal à différentes échelles. Nous avons montré dans des travaux antérieurs, que la transformée en ondelettes donne des résultats intéressants sauf dans certains cas où les singularités sont indiscernables sur les coefficients de la transformée en ondelettes [9]. En effet, le signal de parole présente des singularités lissées aux instants d'ouverture et de fermeture de la glotte [10]. Il est ainsi assez difficile dans certains cas de détecter ces instants sur le signal de parole. Pour circonvenir à ce problème, il est alors judicieux d'opérer une combinaison non linéaire des coefficients de la transformée en ondelettes à différentes échelles afin de donner une meilleure estimation des instants d'excitation de la source.

2. MODULES MAXIMA DE LA TRANSFORMEE EN ONDELETTES

Ce paragraphe montre que les coefficients de la transformée en ondelettes présentent des modules maxima aux singularités du signal. Nous allons d'abord montrer que la transformée en ondelette agit comme un opérateur différentiel multiéchelle d'ordre n, lorsque l'ondelette possède n moments nuls [6]. La transformée en ondelettes est alors la dérivée n^{ième} de la fonction f lissée par une fonction.

La transformée en ondelettes d'une fonction f(t) peut se mettre sous la forme d'un produit de convolution

$wf(u, s) = f * \overline{\psi}_s(u)$, ψ étant l'ondelette mère, avec $\overline{\psi}_s(t)=\frac{1}{\sqrt{s}}\psi\left(\frac{-t}{s}\right)$. Dans laquelle une ondelette à n moments nuls s'écrit comme la dérivée nième d'une fonction de lissage θ ; La transformée en ondelettes s'écrit alors:

$wf(u, s) = s^n \frac{d^n}{du^n}(f * \overline{\theta}_s)(u)$, Le signal est lissé par la fonction $\overline{\theta}_s$.

Comme la transformée en ondelettes de f peut s'écrire comme un opérateur différentiel multiéchelle de f convolué par une fonction de lissage appropriée à l'échelle, lorsque l'ondelette n'a qu'un seul moment nul, les modules maxima de $|Wf(u,s)|$ correspondent aux maximums de la dérivée première de la fonction lissée. Quand la transformée en ondelettes de f n'a pas de maximum local aux fines échelles, alors f est localement régulière. Les singularités se trouvent ainsi au niveau des abscisses où convergent les modules maxima d'ondelettes aux fines échelles, et sont caractérisées par l'ordre de l'ondelette.

3. METHODE DU PRODUIT MULTIECHELLE (MPM)

Le produit des coefficients de la transformée en ondelettes à travers un nombre d'échelle a été fréquemment utilisé en traitement d'image. Pour caractériser les maxima associés aux véritables discontinuities du signal, Mallat et Zhong [11] analysent les propriétes des modules maxima des singularités à travers les échelles. La dérivée de Gaussienne et la spline quadratique sont les ondelettes particulièrement utilisées dans ce cas. La simple multiplication des transformées en ondelettes à des échelles adjacentes, fait ressortir les discontinuités [12], [13], [14]. Sadler et Swami [12] ont étudié l'efficacité de la méthode du produit multiéchelle en présence de bruit. Nous proposons d'utiliser la méthode du produit multiéchelle (MPM). Cet algorithme procède au calcul du produit des coefficients de la transformée en ondelettes pour différentes échelles dyadiques successives. Dans cette opération non linéaire sur f, les singularités produisent à travers les échelles des pics observés sur les coefficients de la transformée en ondelettes qui sont renforcés par le produit multiéchelle. La combinaison non linéaire tend à rehausser les maxima en atténuant les faux pics. Les pics de la transformée vont s'aligner à travers les échelles et non pas pour toutes les échelles parce qu'en augmentant l'effet du lissage, les singularités voisines vont interférer. Ainsi le choix d'une valeur d'échelle assez élevée, fait perdre l'alignement des pics dans le produit multiéchelle. Le nombre impair des termes du produit permet de préserver le signe de la singularité. Trois échelles dyadiques consécutives suffisent généralement pour la détection des pics par le produit.

4. APPLICATION DE MPM SUR LE SIGNAL DE PAROLE

L'ondelette utilisée dans ce travail est la dérivée première de gaussienne. Cette ondelette assure la détection des discontinuités présentes sur le signal de parole; de plus cette ondelette assure la propagation des maxima aux fines échelles [6].

La base de sons de l'Université de Keele est utilisée. Elle est composée de 5 voix féminines et 5 voix masculines. La base contient le signal de parole et le signal EGG. Les deux signaux sont échantillonnés à la cadence 20 kHz. Le signal EGG permet de donner les

instants références des GCIs et GOIs. En effet la dérivée du signal EGG montre deux sortes de pics opposés; les grands pics et qui sont des minimums, concordent à la fermeture de la glotte et les petits pics qui sont des maximums, correspondent à l'ouverture glottique [4].

La figure 1 montre le signal de parole, les coefficients de la transformée en ondelettes à trois échelles, le produit multiéchelle (PM) et enfin la dérivée du signal EGG prise comme référence. Les lignes de maxima de plus forte amplitude correspondent aux instants de fermeture de la glotte repérés sur le signal DEGG. Les GOIs ne sont pas discernables ici sur les transformées en ondelettes.

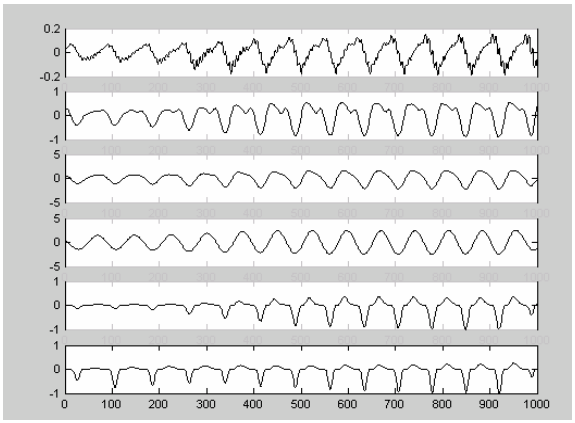


Figure 1 : Signal de parole, locuteur f4, voyelle /o/, mot /north/,3 échelles de transformées en ondelettes, le PM et le DEGG.

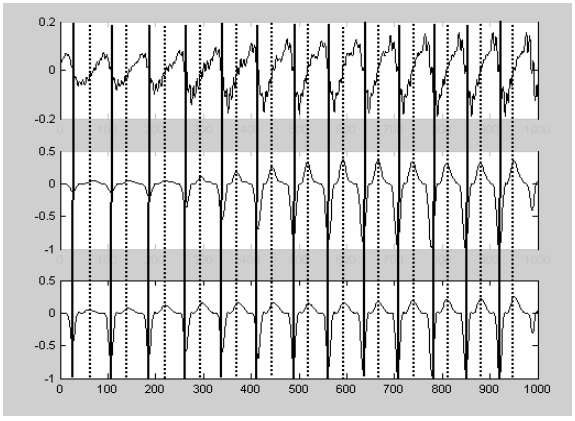


Figure 2 : Signal de parole locuteur f4, PM, DEGG.

La figure 2 montre une tranche d'une zone voisée du locuteur f4 suivie du produit multiéchelle du signal de parole et du signal DEGG. Le produit montre alors deux types d'extrema ; des minimums de forte amplitude marqués par des traits continus et qui correspondent aux instants de fermeture de la glotte et des maximums de moins forte amplitude marqués par des traits interrompus et qui coïncident avec les instants d'ouverture de la glotte donnés par le signal DEGG. Dans ce cas la coïncidence est parfaite.

La figure 3 donne un second exemple : signal de parole, PM et DEGG pour le locuteur f1. Notons également la parfaite correspondance entre les extrema fournis par le produit multiéchelle et ceux donnés par le signal DEGG aussi bien pour l'ouverture que la fermeture de la glotte.

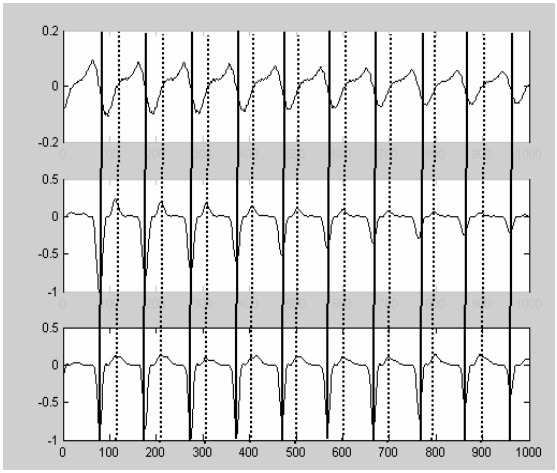


Figure 3 : Signal de parole locuteur f1, PM, DEGG.

Le calcul du produit des coefficients de la transformée en ondelettes du signal de parole avec l'ondelette dérivée première de gaussienne aux trois premières échelles dyadiques, permet une détection précise des GCIs et des GOIs grâce aux pics qui deviennent plus marqués à la fermeture de la glotte. Et l'apparition de pics à l'ouverture. Le signal issu du produit présente une allure semblable au signal DEGG.

5. EVALUATION DE LA METHODE MPM

Cette approche a été évaluée sur la base de sons de l'université de Keele. Nous avons procédé à la comparaison des valeurs instantanées des GCIs et GOIs références donnés par le signal DEGG et des GCIs et GOIs mesurés par la méthode du PM. Les valeurs obtenues dont les erreurs sont inférieures à 0.25 ms par rapport aux valeurs de références sont considérées correctes. L'erreur comprise entre 0.25 ms et 1 ms est comptée comme erreur fine. Les erreurs dépassant 1 ms sont comptées comme grosses. Sur la totalité des mesures des GCIs et GOIs références, nous déterminons le pourcentage des mesures estimées correctes (MC), le pourcentage de mesures correspondant aux erreurs fines (EF), le pourcentage de mesures correspondant aux erreurs grosses (EG), le pourcentage des mesures manquées (MM : cas où le PM ne montre pas un pic au GCI) et le pourcentage de fausses alarmes (FA : cas où le produit montre plus qu'un pic pour un GCI référence donné). L'évaluation est appliquée sur les segments voisés donnés comme tels par la base de sons. Les tableaux 1 et 2 présentent respectivement les performances de la méthode du produit pour la détection des GCIs et des GOIs pour les locuteurs hommes et femmes de la base.

Table 1 : Résultats d'évaluation de la méthode du produit pour la détection des GCIs.

	MC	EF	EG	MM	FA
Hommes	93.73	2.79	1.95	1.51	7.89
Femmes	96.67	1.41	0.22	1.67	0.51
La base	95.70	1.87	0.79	1.62	2.95

Table 2 : Résultats d'évaluation de la méthode du produit pour la détection des GOIs.

	MC	EF	EG	MM	FA
Hommes	49.03	16.36	11.01	23.59	3.38
Femmes	86.12	7.96	1.17	4.76	0.40
La base	73.67	10.78	4.47	11.08	1.4

La méthode proposée est plus performante pour la détection des instants de fermeture de la glotte que pour les instants d'ouverture. Toutefois, elle donne des résultats satisfaisants au regard de l'existant. Les deux tableaux montrent que la méthode proposée est globalement plus performante pour les femmes que pour les hommes aussi bien pour la détection des GCIs que des GOIs.

6. CONCLUSION

La détection des instants de fermeture et d'ouverture de la glotte est opérée à travers l'estimation des discontinuités du signal de parole. Les points caractéristiques du signal de parole sont déterminés par la méthode du produit multiéchelle. Cette méthode consiste à calculer les coefficients de la transformée en ondelettes, le produit de ces coefficients puis détecter les minima et maxima pour localiser respectivement les GCIs et GOIs. Le produit de trois échelles garantit la conservation du signe de la discontinuité. Le produit permet de renforcer les pics aux instants de fermeture de la glotte le long des échelles et de réduire les faux pics particulièrement aux instants d'ouverture glottique. L'évaluation de cette approche sur la base de Keele a donné un taux de mesures correctes de l'ordre de 96 % pour la détection des GCIs et un taux de 74% pour le repérage des GOIs. Ces résultats comptent parmi les meilleurs obtenus à ce jour.

BIBLIOGRAPHIE

[1] B. Yegnanarayana et R. N. J. Veldhuis. Extraction of Vocal-Tract System Characteristics from Speech Signals. *IEEE Trans. Speech and Audio Signals Proc.*, volume 6, pages 313-337, 1998.

[2] J.G. McKenna. Automatic glottal closed-phase location and analysis by Kalman filtering. in Proc. ISCA 2001, Tutorial and Research Workshop on Speech Synthesis, 2001.

[3] C. Hamon, E. Moulines and F. Charpentier. A Diphone Synthesis System Based on Time Domain Prosodic Modifications of Speech. In Proc. *Intl. Conf. Acoustics, Speech and Signal Processing*, pages 238-241, 1989.

[4] N. Henrich, C. d'Alessandro, M. Castellengo et B. Doval. Mesures électroglottographiques du quotient d'ouverture glottique en voix parlée et chantée. In *Proc. 23èmes Journées d'Etude sur la Parole*, 2000.

[5] N. Henrich. *Etude de la Source Glottique en Voix Parlée et Chantée: Modélisation et Estimation, Mesures Acoustiques et Electroglottographiques*. Phd Thesis, Paris 6 University, 2001.

[6] S. Mallat. *Une Exploration des Signaux en Ondelettes*. Les Editions de l'Ecole Polytechnique, Paris, 2000.

[7] L. Janer, J.J Bonet et E.L Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. *In Proc. Intl. Conf. on Spoken Language Processing*, 1996.

[8] S. Kadambe et G.F. Bourdeaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. On Information Theory*, volume 38, 1992.

[9] A.B. Slimane, A. Bouzid et N. Ellouze. Wavelet Decomposition of Voiced Speech and Mathematical Morphology Analysis for Glottal Closure Instants Detection. *In Proc. European Signal Processing Conference*, volume 3, pages 81-84, 2002.

[10] A. Bouzid et N. Ellouze. Caractérisation des Singularités du Signal de Parole. Colloque sur le traitement du signal et des images GRETSI, Tome 1, pages 273-276, 2003.

[11] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 14, pages 710-732, 1992.

[12] B.M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. *IEEE Trans. Information Theory*, vololume 45, pages 1043-1051, 1999.

[13] B.M. Sadler, T. Pham and L.C. Sadler. Optimal and wavelet-based shock wave detection and estimation. *Journal of Acoustical Society of America*, volume 104, pages 955-963, 1998.

[14] P. Bao, L. Zhang and X. Wu. Canny. Edge detection enhancement by scale multiplication. *IEEE Trans. on Pattern Analysis and Machine Intelligenc*e, vol. 27, no. 9, pages 1485-1490, 2005.

Modélisation 2D (« fréquence-temps ») des amplitudes spectrales

Mohammad Firouzmand & Laurent Girin

ICP – INPG/Univ. Stendhal/CNRS
B.P.25-38040 Grenoble France
{girin, firouz}@icp.inpg.fr

Abstract

This paper presents a method for modeling the spectral amplitude parameters of speech signals in "two dimensions" (2D). It consists in two cascaded modeling: the first one along the frequency axis is usual, since it consists in modeling the log-scaled spectral envelope with a sum of Discrete Cosine (DC) functions. The second one, along the time axis, consists in modeling the trajectory of the envelope DC parameters by another similar DC model. An iterative algorithm that optimally fits this 2D-model, taking into account perceptual criterions, is proposed. This approach is shown to provide an efficient representation of speech spectral amplitude parameters in terms of coefficient rates, while providing good signal quality, opening new perspectives in very-low bit-rate speech coding.

1. Introduction

Le modèle sinusoïdal de la parole (SMS) [1] a été largement étudié depuis les années 80 et appliqué avec succès à un grand nombre d'applications, tel que le codage et la transformation de la parole [2-4]. Il consiste à modéliser le signal comme une somme de I sinusoïdes:

$$s(n)=\sum_{i=1}^{I} A_i(n)\cos[\theta_i(n)] \quad \text{avec} \quad \theta_i(n)=\sum_{k=0}^{n}\omega_i(k)+\theta_i(0) \qquad (1)$$

Les paramètres du SMS, amplitudes $A_i(n)$, fréquences $\omega_i(n)$ et phases $\theta_i(n)$, évoluent lentement au cours de temps. Un système d'analyse-synthèse basé sur ce modèle repose sur la mesure « à court terme » (CT) de ces paramètres aux centres de trames d'analyse consécutives, puis l'interpolation CT des valeurs mesurées consécutives afin de reconstruire le signal entier [1]. « Court terme » dénote habituellement une durée d'environ 10-30ms.

Dans une série d'études récentes [5][6], nous avons proposé de modéliser les paramètres du SMS à « long terme » (LT). Ceci signifie que les trajectoires temporelles de ces paramètres ont été modélisées sur de longues sections de parole, au-delà de la longueur de trame usuelle en analyse-synthèse à court terme. Par exemple, dans ces études, un modèle LT unique a été employé pour coder chaque trajectoire des paramètres sur chaque section de parole entièrement voisée. Une telle approche a été appliquée sur les trajectoires d'amplitude [5][6] et de phase [6] en utilisant un modèle en Cosinus Discrets (MCD) comme modèle LT. Un algorithme itératif incluant des contraintes perceptives a permis d'estimer conjointement l'ordre optimal du modèle et ses coefficients. Le nombre de ces coefficients a été réduit significativement par rapport à la modélisation CT, ce qui ouvre une nouvelle voie pour le codage de parole/audio à très bas débit. Dans ce papier, nous étendons et raffinons cette approche en ajoutant une nouvelle étape de modélisation le long de l'axe des fréquences avant de considérer l'axe du temps : nous modélisons d'abord l'enveloppe des amplitudes par un modèle en Cosinus Discrets, comme proposé dans [7], mais avec une version raffinée qui tient compte des contraintes perceptives, inspirée de [5]. Puis, nous appliquons un deuxième MCD sur l'axe du temps pour modéliser à long terme la trajectoire des coefficients résultant de la modélisation d'enveloppe. La modélisation de l'enveloppe spectrale effectuée avant la modélisation à LT permet : 1) de résoudre le problème de « taille variable » des jeux de paramètres d'amplitudes d'une trame d'analyse à l'autre, dû aux variations de la fréquence fondamentale (et à la présence de composantes de bruit), car l'enveloppe est modélisée en utilisant un ordre fixe sur la section de parole considérée à LT ; 2) de réduire la taille de ces jeux de paramètres avant la modélisation à LT, puisque l'ordre du modèle d'enveloppe est généralement très inférieur au nombre d'amplitudes mesurées ; c'est un point important dans l'optique de l'utilisation de cette modélisation 2D pour le codage de parole à très bas débit ; et 3) de s'adapter aux transformations du signal en fréquence, telle que le *pitch-scaling* par exemple. Notons que dans cet article, nous considérons seulement la modélisation 2D des paramètres d'amplitude. Comme un des buts sous-jacents de cette étude est de fournir des outils efficaces pour un codeur à très bas débit, l'information de phase est réduite à la trajectoire de fréquence fondamentale pour les sections considérées qui sont toutes voisées. La modélisation à long terme de la trajectoire de fréquence fondamentale peut être réalisée en parallèle, comme dans [6].

Ce papier est organisé comme suit. Le processus complet d'analyse-modélisation-synthèse est décrit dans la Section 2, incluant la description du MCD, l'approche 2D, et l'algorithme d'ajustement. Des résultats sont donnés dans la Section 3.

2. Modélisation 2D des amplitudes

Nous supposons que le signal de parole est d'abord segmenté en parties voisées et non voisées par des classificateurs habituels (non décrits ici), et nous considérons ici le problème de modélisation en 2D des paramètres d'amplitude d'une section voisée entière $s(n)$, où $n = 0$ à N. Nous présentons d'abord le processus d'analyse des amplitudes, puis le modèle 2D et ensuite l'algorithme qui est utilisé pour adapter le modèle aux mesures d'amplitude.

2.1. Analyse

La première étape du processus est d'extraire les jeux successifs de paramètres d'amplitude devant être modélisés en 2D. Bien que le processus de modélisation 2D est un processus à LT selon l'axe du temps, l'analyse de chaque jeu d'amplitude est fournie sur une base habituelle à court terme.

Les expériences décrites dans cet article ont été conduites avec une analyse pitch-synchrone, pour être cohérentes avec nos études précédentes [5][6], mais des techniques d'analyse plus habituelles, par exemple basées sur la Transformée de Fourier à court terme, peuvent être utilisées. Ainsi, les signaux sont d'abord « pitch-marqués » en employant le logiciel Praat [8]. Cela signifie que les frontières des quasi-périodes du signal ont été automatiquement détectées et ces quasi-périodes sont utilisées comme trame d'analyse. La fréquence fondamentale ω_0^k est alors directement donnée par l'inverse de la période. Puis, étant donnée ω_0^k, les I_k amplitudes A_i^k correspondant à chaque harmonique à la fréquence $i\,\omega_0^k$ et mesurées au centre n_k de chaque période, ont été estimées en utilisant le procédé employé par George et Smith dans [4]. Pour la suite, il est important de noter que I_k dépend de ω_0^k. L'estimation des amplitudes est basée sur la minimisation de l'erreur entre le modèle sinusoïdal harmonique et le signal selon un critère des moindres carrés (MMSE). Cette procédure fournit une estimation des paramètres très précise avec un coût de calcul très bas.

2.2. Modélisation de l'enveloppe spectrale

Dans nos études précédentes, nous avons considéré la modélisation des amplitudes directement selon l'axe du temps [5]. Donc, à la fin du processus d'analyse, les amplitudes étaient réordonnées comme I jeux de K amplitudes (^{t} dénote le vecteur/matrice transposé):

$$\boldsymbol{A}_i = \left[A_i^1 \quad A_i^2 \quad \ldots \quad A_i^K\right]^t, \quad i = 1 \text{ à } I,$$

Et la trajectoire résultante était modélisée par un MCD. Alternativement, dans la présente étude, les amplitudes sont d'abord ordonnées comme couramment selon l'axe des fréquences, comme K jeux de I_k valeurs :

$$\boldsymbol{A}^k = \left[A_1^k \quad A_2^k \quad \ldots \quad A_{I_k}^k\right]^t, \quad k = 1 \text{ à } K$$

Chaque jeu de paramètres est alors remplacé par un modèle d'enveloppe spectral. Nous employons le MCD de [8] qui consiste à modéliser l'enveloppe spectrale (en échelle log) par une somme de fonctions cosinus :

$$A_{DCM}^k(\omega) = d_0^k + 2\sum_{m=1}^{M} d_m^k \cos(2\pi m\,\omega) \qquad (2)$$

où M est un nombre entier positif qui est l'ordre du modèle. Bien que le nombre d'harmoniques puisse varier d'une trame à l'autre, M a une valeur fixe pour toutes les trames d'une section voisée modélisée à LT. Nous verrons dans la Section 2.4 comment M est estimé pour chaque section. Une fois M estimé, le vecteur $\boldsymbol{D}^k$ des coefficients du modèle est estimé par une version pondérée de la procédure MMSE d'ajustement du MCD avec le jeu d'amplitudes mesurées (en échelle log), minimisant l'erreur :

$$\varepsilon_k = \sum_{i=1}^{I_k} w_i^k \left\| A_i^k - A_{DCM}^k(i\omega_0^k) \right\|^2 \qquad (3)$$

Si on dénote par $\boldsymbol{M}_k$ la matrice $I_k \times (M+1)$ de terme général $\cos(2\pi m i \omega_0^k)$ (avec un facteur 2 pour les colonnes $m>1$), et si on dénote par $\boldsymbol{W}_k$ la matrice diagonale qui contient les poids de (3) mis au carré sur sa diagonale (ce poinds sont estimés à partir de contraintes perceptives dans l'algorithme de la Section 2.4), on a :

$$\boldsymbol{D}^k = \left(\boldsymbol{M}_k^t W_k \boldsymbol{M}_k\right)^{-1} \boldsymbol{M}_k^t W_k A^k \qquad (4)$$

2.3. Modélisation des trajectoires d'enveloppe

Une fois que la modélisation spectrale a été faite pour toutes les trames de la séquence de parole considérée (section voisée), la seconde étape de modélisation est la modélisation de la trajectoire temporelle des coefficients d_m^k du modèle d'enveloppe. Ainsi, ces coefficients sont d'abord ré-ordonnés le long de l'axe du temps comme M+1 jeux de K-vecteurs (comme fait directement pour les amplitudes dans [4]):

$$\boldsymbol{D}_m = \left[d_m^1 \quad d_m^2 \quad \ldots \quad d_m^K\right]^t \quad \text{pour } m = 0 \text{ à } M$$

Un deuxième modèle MCD est alors appliqué sur chacune des M+1 trajectoires, selon la même approche que celle utilisée dans [5] pour modéliser les amplitudes à LT :

$$d_m^{DCM}(n) = c_0^m + 2\sum_{p=0}^{P_m} c_p^m \cos\left(2\pi p \frac{n}{2N}\right) \quad \text{pour } m = 0 \text{ à } M \qquad (5)$$

Comme pour la modélisation d'enveloppe, les coefficients du modèle sont estimés par minimisation des moindres carrés pondérés (rappelons que les index n_k sont les centres des K trames d'analyse) :

$$\varepsilon_m = \sum_{k=1}^{K} w_m^k \left\| d_m^k - d_m^{DCM}(n_k) \right\|^2 \qquad (6)$$

Si $\boldsymbol{M}_m$ dénote ici la matrice $K \times (P_m+1)$ de terme général $m_{kp} = \cos(2\pi p n_k / 2N)$, et si $\boldsymbol{W}_m$ dénote la matrice diagonale qui contient sur sa diagonale les poids de (6) mis au carré, le vecteur des coefficients du modèle est donné de façon similaire à (4) par :

$$\boldsymbol{C}^m = \left(\boldsymbol{M}_m^t \boldsymbol{W}_m \boldsymbol{M}_m\right)^{-1} \boldsymbol{M}_m^t \boldsymbol{W}_m \boldsymbol{D}_m \qquad (7)$$

Cependant, dans ce cas, l'ordre du modèle P_m dépend de la section de parole considérée et potentiellement du rang m du coefficient de l'enveloppe. En effet, l'évolution de l'enveloppe spectrale peut varier considérablement, par exemple selon la longueur de la section considérée, la séquence de phonèmes, le locuteur, la prosodie, etc. Nous présentons ci-dessous l'algorithme que nous utilisons pour estimer conjointement l'ordre P_m, les poids qui sont utilisés dans (4) et (7), et également l'ordre M du modèle d'enveloppe. Notons que le modèle 2D proposé peut être efficacement exploité pour le codage de parole à très bas débit si dans la pratique on constate que l'ordre estimé P_m est significativement inférieur à K. Ce point sera discuté plus en détail dans la Section 4.

2.4. Estimation de l'ordre et algorithme d'adaptation

Par simplicité, dans cette étude, nous prenons le même ordre $P_m = P$ pour tous les vecteurs $\boldsymbol{D}_m$, $m = 0$ à M. L'algorithme ci-dessous peut être raffiné pour garder la possibilité d'un ordre spécifique pour chaque rang de coefficient. L'algorithme est divisé en deux parties: La première partie consiste à régler M de façon à conjointement adapter K modèles d'enveloppe optimaux aux K jeux d'amplitudes mesurées selon un critère perceptuel moyen. Ces modèles d'enveloppe seront utilisés comme une référence pour la deuxième partie de l'algorithme qui traite de la dimension temporelle, incluant une estimation itérative de l'ordre P.

Première partie de l'algorithme : modélisation des enveloppes (*M* est initialisé à une valeur arbitraire, typiquement 10)

1) Pour chaque index du temps *k*, *k* = 1 à *K*, calculer le seuil de masquage fréquentiel global $T^k(\omega)$ associé au vecteur d'amplitude $\boldsymbol{A}^k$ en employant le modèle de [9].

2) Initialiser *K* vecteurs de poids $\boldsymbol{W}_k$, chacun de longueur I_k, avec tous les éléments à un. Itérer alors le processus suivant de l'étape 3 à l'étape 5, jusqu'à ce que chaque ratio R^k de l'étape 5 soit maximisé.

3) Calculer *K* vecteurs MCD d'enveloppe avec (4).

4) Pour chaque trame *k*, augmenter les poids où l'erreur de modélisation dépasse le seuil masquage, selon (*square* et *max* dénotent les fonctions carré et maximum (élément-par-élément), *diag* dénote la fonction qui produit une matrice diagonale à partir d'un vecteur, les éléments du vecteur étant mis sur la diagonale) :

$$\boldsymbol{E}_i^k = \frac{1}{2} square(\boldsymbol{A}_i^k - \boldsymbol{M}_k \boldsymbol{D}^k)$$

$$d\boldsymbol{W}_k = diag(max(\boldsymbol{E}_i^k - \boldsymbol{T}^k(i\omega_0^k),0))$$

$$\boldsymbol{W}_k = \boldsymbol{W}_k + d\boldsymbol{W}_k / max(d\boldsymbol{W}_k)$$

5) Calculer le pourcentage R^k des éléments nuls de $d\boldsymbol{W}_k$.

6) Une fois que tous les R^k sont maximisés, calculer la valeur moyenne R_{min} de ces rapports. Si R_{min} est supérieur à un rapport prédéfini R_{target} (en général 0,8 à 0,9), *M* est diminué de 1, sinon, *M* est augmenté de 1. Retourner ensuite à l'étape 2 jusqu'à stabilisation de *M*.

Deuxième partie de l'algorithme : modélisation des enveloppes au cours de temps

7) Initialiser *P* à une valeur arbitraire significativement inférieure à *K*, par exemple, la partie entière de *K*/4.

8) Pour *m*=0 à *M*, calculer le modèle LT (2ème MCD) de trajectoire des coefficients d'enveloppe (1er MCD) avec (7).

9) Décoder les *K* modèles d'enveloppe à partir des *M*+1 MCD à long terme par : $\hat{\boldsymbol{D}}_m = \boldsymbol{M}_m \boldsymbol{C}^m$ (équivalente à (5) appliquée aux instants n_k).

10) Réordonner les coefficients d'enveloppe décodés selon l'axe des fréquences et décoder les amplitudes par : $\hat{A}^k = \boldsymbol{M}_k \hat{\boldsymbol{D}}^k$ (équivalent à (2) appliquée aux fréquences harmoniques et avec les coefficients d'enveloppe décodés).

11) Calculer le rapport R_{min} de l'étape 6 en remplaçant les amplitudes modélisées à l'étape 6 (modélisation « 1D ») par celles issues de l'étape 10 (modélisation « 2D »). Si R_{min} dépasse un rapport prédéfini (typiquement $0.75 \times R_{target}$), *P* est diminué de 1, sinon *P* est augmenté de 1. Retourner ensuite à l'étape 8 jusqu'à stabilisation de *P*.

Notons que dans cette étude, tous les poids fonctions du temps sont mis à un, *i.e.*, aucune pondération n'est en réalité effectuée le long de l'axe de temps dans la deuxième partie de l'algorithme. Un ajustement itératif des poids semblable à ce qui a été fait dans la première partie de l'algorithme et dans nos études précédentes [5][6] pourrait probablement aboutir à diminuer encore *P*. Cependant, dans la plupart des expériences que nous avons faites, ce raffinement augmente considérablement la complexité de calcul pour une amélioration faible (dans la plupart des cas, la valeur optimale de *P* est obtenue à la première itération sur les poids). De nouveaux travaux doivent être menés concernant ce point.

2.5. Synthèse

La synthèse est réalisée en appliquant d'abord l'interpolation linéaire entre les amplitudes (ramenées à l'échelle linéaire) résultant de l'algorithme ci-dessus. Cette interpolation inclut le processus habituel de « naissance et de mort » pour les harmoniques qui dépassent la fréquence de Nyquist [1]. Notons qu'un codeur de parole à bas débit utilisant la méthode proposée doit coder la trajectoire de fréquence fondamentale qui est employée dans l'étape 10. L'équation (1) est ensuite utilisée pour produire le signal de synthèse (fréquences et phases mesurées sont interpolées en utilisant la procédure classique de [1], puisque dans cette étude on s'intéresse seulement à la modélisation 2D des amplitudes). Dans cette étude, le processus d'analyse-modélisation-synthèse concerne seulement les parties voisées de parole. Les sections non voisées sont conservées telles quelles et sont concaténées avec les sections voisées modélisées avec une pondération locale pour éviter des artefacts audibles [4].

3. Expériences

Dans cette section, nous décrivons un ensemble d'expériences qui ont été conduites pour évaluer la modélisation des paramètres d'amplitudes par le modèle 2D présenté. Nous avons utilisé des signaux de parole échantillonnés à 8 kHz et produits par 12 locuteurs (six masculins et six féminins). Un total d'environ 3500 segments voisés de différentes tailles ont été modélisés, représentant plus de 13 minutes de parole.

3.1. Précision de la modélisation

D'abord, l'algorithme s'adapte généralement correctement aux différentes configurations des sections modélisées. Par exemple, l'ordre *M* peut varier avec des petites valeurs (par exemple 4) pour des spectres assez « pauvres », des valeurs habituelles pour coder la parole féminine (en général 10-11) et masculine (typiquement 15-16, voir [7]), et plus pour des spectres « riches ». Le long de l'axe de temps, l'ordre *P* varie également beaucoup, selon la longueur et le contenu de la section de parole modélisée. Généralement, le modèle 2D fournit des valeurs d'amplitudes qui sont assez proches des amplitudes mesurées. Ceci est garanti par la contrainte perceptuelle qui guide le comportement de l'algorithme d'ajustement: à la fin de l'algorithme, $0.75 \times R_{target}$ pourcents des amplitudes modélisées sont assurées de vérifier cette contrainte (l'erreur de modélisation est au-dessous du modèle de seuil de masquage, et on s'attend ainsi à ce qu'elle soit inaudible). Le réglage de R_{target} à 0,75 assure qu'au moins la moitié des amplitudes sont correctement modélisés selon la contrainte perceptuelle. Les expériences ont montré qu'il n'est pas très efficace d'essayer d'augmenter ce rapport global. En effet, étonnamment, la deuxième étape de modélisation (au cours de temps) fournit généralement des trajectoires d'amplitudes assez « bruitées », qui aboutissent à une diminution de la qualité du signal de synthèse. Les trajectoires modélisées par le MCD à long terme ont bien la propriété intrinsèque d'être lisses, puisque le modèle est composé de fonctions de type cosinus, elles-mêmes lisses. Cependant, des trajectoires lisses de coefficients d'enveloppe ne fournissent

pas nécessairement une suite d'enveloppes spectrales évoluant de façon régulière (et ainsi pour les amplitudes). En d'autres termes, des variations régulières des coefficients d'enveloppe entre deux trames consécutives peuvent produire des discontinuités gênantes pour les trajectoires d'amplitudes. Par conséquent, plutôt que de modifier profondément la méthode proposée, un post-filtrage est appliqué sur les trajectoires d'amplitudes pour les lisser. Un filtre médian simple s'est montré efficace pour régulariser les trajectoires et permettre la synthèse de signaux de bonne qualité (voir la Figure 1).

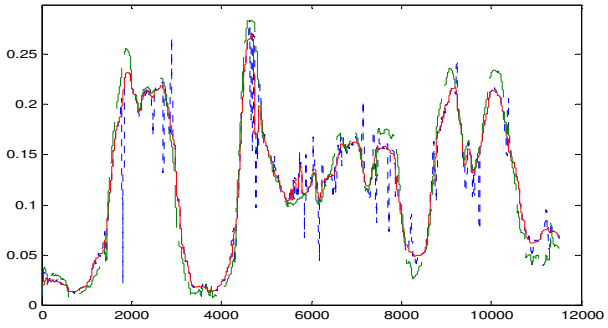


Figure 1 : Trajectoire des amplitudes mesurées (tirets verts), modélisées en 2D avant (pointillés bleus) et après (continu rouge) post-filtrage, pour la 1ère harmonique d'une longue séquence voisée de voix de femme (1,33s à 8 kHz, K=408).

3.2. Débit des coefficients

Dans cette section, nous donnons des exemples de débits moyens pour les coefficients du modèle 2D (*i.e.*, ceux de $\boldsymbol{C}^m$, soit l'information à transmettre dans un système de communication pour coder les amplitudes spectrales avec la technique proposée). Ces débits ont été calculés sur les 13 minutes de parole testées. Avec le rapport-cible global réglé à 90%, nous obtenons un débit moyen de 341 coefficients/s pour les voix féminines, et 404 coefficients/s pour les voix masculines. Ces débits peuvent être comparés à ceux d'une approche habituelle à court terme avec une fenêtre de 20ms : si on suppose que des spectres de voix de femme sont codés avec 11 coefficients MCD en moyenne (pour 8 kHz de parole), et que ceux de voix d'homme exigent 17 de ces coefficients, nous obtenons respectivement 11×50 = 550 coefficients/s et 17×50 = 850 coefficients/s. En faisant la moyenne, nous obtenons environ 370 coefficients/s pour le modèle 2D contre 700 coefficients/s pour l'approche habituelle « 1-D ». Ainsi, la stratégie de modélisation à long terme permet de diminuer de façon significative le débit des coefficients (d'un facteur 2 dans cette expérience). Notons que, dans ces expériences, les signaux ont été modélisés avec les deux approches 1D et 2D, et les débits mentionnés ci-dessus fournissent généralement des signaux de qualité semblable (voir ci-dessous).

3.3. Qualité du signal

Deux sujets avec une audition normale ont intensivement écouté les signaux synthétisés. Comme mentionné avant, quand le modèle 2D est appliqué directement, la qualité du signal synthétisé est diminuée par le bruit de modélisation sur les trajectoires d'amplitude. Cependant, l'utilisation d'un filtre médian permet de résoudre ce problème de discontinuité des amplitudes et la qualité du signal synthétisé est alors fortement améliorée sans exiger d'information supplémentaire. Pour des valeurs du rapport-cible d'environ 50%, le signal synthétisé est de bonne qualité, et assez proche de l'original. Si le rapport diminue, les trajectoires des coefficients du MCD à LT sont « simplifiées » (puisque la contrainte sur l'ordre P est relachée), et il en est de même pour les trajectoires de l'enveloppe spectrale (après post-traitement pour le lissage additionnel). Ainsi, le signal restitué, bien qu'ayant une bonne sonorité, s'éloigne du signal original : il tend vers une version hypo-articulée de ce dernier. Ceci confirme l'importance de considérer l'aspect dynamique (temporel) dans la modélisation du signal de parole.

4. Conclusion

Ce travail a confirmé la robustesse et la généralité du modèle en Cosinus Discrets, adéquat pour modéliser l'enveloppe spectrale (comme déjà montré dans [7]) et la trajectoire temporelle de paramètres (comme déjà montré dans [5]). Réunir ces deux aspects dans un modèle 2D a abouti à de nouvelles avancées. Dans cette étude, la raison principale de cette efficacité est la variabilité intrinsèque du débit : l'ordre M du modèle d'enveloppe et l'ordre P du modèle temporel sont tous les deux ajustés sur les caractéristiques locales du signal. L'approche 2D et l'algorithme associé peuvent permettre de réduire significativement le nombre de coefficients pour la représentation des amplitudes spectrales, tout en préservant une qualité raisonnable pour les signaux synthétisés. Les travaux futurs concerneront l'utilisation de l'approche proposée dans un codeur de parole à très bas débit sans contraintes sur le délai de codage-décodage.

5. Références

1. R. J. McAulay & T. F. Quatieri, Speech analysis/ synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, **34**(4), 1986, pp. 744-754.
2. T. F. Quatieri & R. J. McAulay, Shape invariant time-scale and pitch modification of speech, *IEEE Trans. Signal Proc.*, **40**(3), 1992, pp. 497-510.
3. R. J. McAulay & T. F. Quatieri, Sinusoidal coding, in *Speech coding and synthesis*, (W. B. Kleijn & K. K. Paliwal, eds), ch. 4, Elsevier, 1995.
4. E. B. George & M. J. T. Smith, Speech analysis/ synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.,* **5**(5), 1997, pp. 389-406.
5. M. Firouzmand & L. Girin, Perceptually weighted long-term modeling of sinusoidal speech amplitude trajectories, *Proc. IEEE Int. Conf. on Acoustics, Speech & Signal Proc. (ICASSP 2005),* Philadelphia, USA, 2005.
6. L. Girin, M. Firouzmand & S. Marchand, Perceptual long-term variable-rate sinusoidal modeling of speech, *submitted to IEEE Trans. Speech and Audio Proc.*, 2005.
7. O. Cappé, J. Laroche & E. Moulines, Regularized estimation of cepstrum envelope from discrete frequency points, *Proc. IEEE Workshop Applications Signal Proc. Audio Acoustics (WASPAA)*, 1995.
8. www.praat.org
9. ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s, Part 3: Audio, 1992.

Réduction du débit des LSF's par un système d'énumération en treillis.

Bachir BOUDRAA, Malika BOUDRAA, Mouloud DJAMAH, Merouane BOUZID

USTHB, LCPTS, *Faculté d'Electronique et d'Informatique,* BP 32, El alia, Alger, ALGERIE.
b.boudraa@yahoo.fr; mboudraa@usthb.dz.

Bernard GUERIN

ICP-INPG, 46, Avenue Félix Viallet, 38031, Grenoble Cedex, France

ABSTRACT

In the present study, we were interested in the reduction of the bit-rate observed in the speech coder named CELP FS1016 (federal standard developed by the US department of the defense "DoD "). More precisely, the quantization of the Line Spectrum Frequencies (LSF) parameters was concerned. In the standard CELP FS1016, these coefficients are coded with 34 bits. We propose the use of an enumeration technique in conjunction with a treillis search coding schemes that exploits the natural ordering of the LSF. The technique allows reducing the bit rate of the LSF coefficients to 30 bits without decreasing the performance of the coder.

1. INTRODUCTION

Actuellement, la représentation des coefficients LPC par les paires des raies spectrales est très utilisée [1]-[4], car celles-ci possèdent des propriétés naturelles désirables pour une quantification, comme nous allons le préciser dans la section suivante. Plusieurs schémas de quantification des LSF, aussi bien scalaires (QS) que vectorielles (QV), sont rencontrés dans la littérature [1]-[4]. Dans ce travail, nous avons opté pour la QS à cause de la complexité moindre qu'elle présente comparativement à la QV, ce qui était notre exigence de départ, consistant à mettre au point un codeur de parole fonctionnant en full duplex sur DSP TMS 320C25. D'où notre intérêt pour le codeur CELP (Code Excited Linear Prediction) de la norme FS1016 [5], standard développé par le département de la défense des Etats Unis d'Amérique "DoD", qui répondait a priori à notre exigence. Ce dernier utilise 10 coefficients LSF pour représenter l'enveloppe spectrale. Ceux-ci sont quantifiés de façon scalaire et emploient 34 bits/trame, répartis respectivement comme suit : 3, 4, 4, 4, 4, 3, 3, 3, 3, 3 [5]. Dix tables de quantification formant un total de 112 éléments sont utilisées. Une fine analyse de ces tables révélera que le FS1016 utilise une numérotation dont certaines valeurs ne peuvent être retenues car introduisant des combinaisons de coefficients LSF qui rendraient le filtre de synthèse instable. En effet, les 34 bits précédents peuvent donner $N=\prod_{i=1}^{10} 2^{b_i}$, soit 17179869184 combinaisons possibles où b_i serait le nombre de bits accordé pour pouvoir indexer tous les nivaux de quantification de la $i^{ème}$ table. Cependant, l'ordonnancement naturel des LSF va nous permettre de réduire ce nombre à seulement un maximum de 554958388 (10000100010011111111000110100 en binaire) combinaisons acceptables (soit environ 3.25% du nombre total). Ce chiffre n'exige que 30 bits pour sa représentation.

Tout en gardant une complexité raisonnable, nous avons pu réduire ce débit par l'application d'un système d'énumération en treillis sur les 10 tables précédentes.

La quantification par treillis a eu cette appellation suite à son schéma utilisant des chemins sous forme de treillis. Ces chemins passent d'un niveau de reproduction à un autre se trouvant dans des tables différentes. Plusieurs travaux [6],[7] se sont intéressés à l'application de ce type de quantification aux coefficients LSF. Dans notre cas, nos avons adapté l'algorithme de Malone et Fisher [7] qui nous a permis de récupérer 4 bits/trame ce qui équivaut à un gain de 133 bits/sec.

2. DÉFINITION ET PROPRIETES DES LSF

Les coefficients LSF sont une autre représentation des coefficients de prédiction a_k . Quelques propriétés de ces coefficients LSF peuvent être résumées ainsi [8] :

- Une condition nécessaire et suffisante pour la stabilité du filtre de synthèse *1/A(z)* est que les coefficients LSF doivent respecter la condition suivante :

$$0 < LSF_1 < LSF_2 < \ldots < LSF_p < 0.5 \qquad (1)$$

0.5 étant la fréquence de Nyquist. Cette propriété exprime l'ordonnancement de ces coefficients.

- En général, la sensibilité spectrale de chaque coefficient LSF est localisée. Cela veut dire qu'un changement dans un coefficient causera un changement dans le spectre de puissance près de son voisinage.

- Les LSF sont une représentation fréquentielle dont la quantification peut aisément incorporer des traits très importants pour la perception et la sensibilité de l'oreille humaine en particulier.

3. QUANTIFICATION PAR ENUMERATION EN TREILLIS

Une analyse des combinaisons des différents éléments des tables de la QS utilisée dans la norme FS1016 montre qu'il y a un nombre très important de vecteurs LSF qui, s'ils étaient utilisés, rendraient le filtre de synthèse

instable. Aussi, ce système est conçu pour éviter les combinaisons donnant des vecteurs non acceptables en sortie du quantificateur (soit par exemple $LSF_2 < LSF_1$).
Dans ce qui va suivre nous proposons d'appliquer une méthode d'énumération utilisant un système en treillis réduisant le nombre de bits nécessaires aux coefficients LSF à 30 bits au lieu de 34 bits, sans changer les tables de la quantification scalaire suscitée.

4. ALGORITHME DE CODAGE

Pour comprendre l'algorithme de quantification suivant, on précise les notations suivantes :

- LSF_i est le coefficient d'ordre i obtenu après analyse.
- Le mot-code du coefficient LSF_i est donné par $L\hat{S}F_{i,j}$.

  L'opération de quantification sera notée :

  $QS_{34}(LSF_i) = L\hat{S}F_{i,j}$
- Pour un ordre de prédiction p, le nombre maximal de combinaisons que les tables de cette QS pourraient donner est : $N_{max} = \prod_{i=1}^{p} 2^{b_i}$ où b_i est le nombre de bits accordé à là i^{eme} table.
- $N_{i+1}(j)$ est le nombre de chemins possibles allant du mot-code du coefficient LSF_i dans la table C_i à l'indice j (soit $L\hat{S}F_{i,j}$), pour atteindre le mot-code du dernier coefficient de la table C_p, tout en vérifiant la propriété d'ordonnancement donnée par l'équation (1).

Ainsi, pour une position j dans la table C_{p-1}, le nombre de combinaisons possibles pour atteindre le dernier coefficient dans la table C_p est donné par :

$$N_p(j) = \sum_{\substack{k=1 \\ L\hat{S}F_{p,k} > L\hat{S}F_{p-1,j}}}^{2^{bp}} 1 \quad , \quad 1 \le j \le 2^{b_{p-1}} \qquad (2)$$

Dans le tableau 1, on donne le nombre de tous les chemins possibles menant de n'importe quelle position dans une table C_i quelconque jusqu'à la table C_{10} en respectant la propriété d'ordonnancement.

Pour chaque $L\hat{S}F_{i,j}$, $1 \le i \le p\text{-}1$, $1 \le j \le 2^{b1}$ on a :

$$N_{i+1}(j) = \sum_{L\hat{S}F_{i+1,k} > L\hat{S}F_{i,j}} N_{i+2}(k) \qquad (3)$$

Autrement dit, d'un niveau j dans la table C_i, on aura N_{i+1} chemins possibles assurant une bonne succession des coefficients *LSF*. Ces chemins peuvent se déduire de la connaissance du nombre de chemin à l'ordre $i+2$. Notons que la progression est du type arrière (ou *backward)*.

- Le total des combinaisons possibles est :

$$N(p, \overline{C}) = \sum_{j=1}^{2^{b1}} N_2(j) \qquad (4)$$

$$Où\ \overline{C} = \{C_1, C_2, ..., C_p\}$$

$N_2(j)$ est le nombre de combinaisons possibles vérifiant :

$$L\hat{S}F_{1,j} < L\hat{S}F_{2,k}, \quad 1 \le j \le 2^{b1}, 1 \le k \le 2^{b2}.$$

Formellement, le nombre de bits nécessaires est :

$$K = \lceil log_2(N(p, \overline{C})) \rceil \qquad (5)$$

Dans le cas du codeur CELP FS1016, $N(p, \overline{C})$ vaut 554958388, où $\lceil x \rceil$ désigne le plus petit entier supérieur à x. K vaut 30 bits dans le cas précédent.

- Dans l'algorithme, m désigne le plus petit ordre dans la table C_i tel que le mot-code d'ordre m soit directement supérieur au coefficient LSF_{i-1} dont l'ordre est j dans C_{i-1}, c-à-d : $L\hat{S}F_{i-1,j} < L\hat{S}F_{i,m}$. Par ailleurs, les valeurs N_i sont celles données au tableau. On notera par c le mot-code de l'énumération en treillis qui sera initialisé à zéro.

Tableau 1 : Nombre de chemins possibles N_i, partant de la table C_1 jusqu'à la table C_{10}

	N_i : Nombres de chemin possibles partant de la table C_{i-1}									
Indice dans C_{i-1}	***N2***	***N3***	***N4***	***N5***	***N6***	***N7***	***N8***	***N9***	***N10***	***N11***
1	*97762793*	*8512383*	*896795*	*96022*	*9455*	*1904*	*315*	*52*	*8*	*1*
2	*97762793*	*8512383*	*896795*	*96022*	*9455*	*1904*	*315*	*52*	*8*	*1*
3	*89250410*	*8512383*	*896795*	*96022*	*9455*	*1904*	*315*	*52*	*8*	*1*
4	*80738027*	*8512383*	*896795*	*96022*	*9455*	*1589*	*315*	*52*	*8*	*1*
5	*72225644*	*8512383*	*896795*	*96022*	*9455*	*959*	*263*	*44*	*7*	*1*
6	*55200878*	*8512383*	*800773*	*96022*	*9455*	*644*	*211*	*36*	*6*	*1*
7	*38176112*	*8512383*	*704751*	*77112*	*9455*	*381*	*107*	*20*	*4*	*1*
8	*23841731*	*7615588*	*608729*	*67657*	*7551*	*170*	*63*	*7*	*3*	*1*
9	***554958388***	*6718793*	*512707*	*58202*	*5647*	***9455***	***1904***	***315***	***52***	***8***
10		*5821998*	*416685*	*39292*	*5647*					
11		*4925203*	*320663*	*29837*	*3743*					
12		*4028408*	*243551*	*22286*	*2154*					
13		*3227635*	*175894*	*10992*	*2154*					
14		*2522884*	*117692*	*7249*	*1195*					
15		*1914155*	*78400*	*5095*	*1195*					
16		*1401448*	*48563*	*2941*	*551*					
		97762793	***8512383***	***896795***	***96022***					

Algorithme:

Etape 1 : *Trouver j tel que* $QS_{34}(LSF_1) = L\hat{S}F_{1,j}$ *avec* $1 \le j \le 2^{b1}$ *et mettre* $j^* = j\text{-}1$

Etape 2 : *Initialiser* $c = 0$ *si* $j^* = 1$ *sinon*

$$c = \sum_{k=1}^{j^*} N_2(k)$$

Etape 3 : *Pour i = 2 à p*

$m = arg\ min_{k=1..2^{bi}} (L\hat{S}F_{i-1,j^*} < L\hat{S}F_{i,k})$

Trouver j tel que $QS_{34}(LSF_i) = L\hat{S}F_{i,j}$

avec $1 \le j \le 2^{bi}$ *et mettre* $j^* = j\text{-}1$

si $j^* \neq m$ *calculer* $c = c + \sum_{k=m}^{j^*} N_{i+1}(k-1)$

Etape 4 : *Le mot-code d'énumération en treillis sera c.*

Pour une meilleure illustration de l'algorithme, nous prenons l'exemple de la quantification du vecteur LSF suivant :

LSF = [0.0409 0.0869 0.1312 0.1621 0.2340 0.2649 0.3052 0.3321 0.3830 0.4073],

obtenu après analyse d'une trame de 30 ms du signal parole. Le déroulement de l'opération de quantification suit les étapes suivantes : dans les deux premières étapes, on cherche le mot-code du premier coefficient LSF_1, en appliquant l'algorithme de la quantification scalaire du FS1016. On s'intéresse ici à son indice. La table C1 de la norme donne un indice *j=6* et par conséquent on aura : $j^*=6\text{-}1=5$. Comme $j^* \neq 1$, on calculera *c* en utilisant le tableau 1.

$$c = \sum_{k=1}^{5} N_2(k) = 437739667 \quad (6)$$

Dans l'étape 3, on déroule une boucle qui commence de LSF_2 pour atteindre LSF_{10}. Pendant ce déroulement, on détermine d'abord l'ordre *m* comme indiqué plus haut. Dans le cas de LSF_2, *m* vaut *6.* Par la suite, on cherche, comme dans la première étape, l'indice *j* du représentant du coefficient LSF_2. Dans cet exemple *j* vaut *13* et on aura alors $j^* = 12$.

Comme $j^* \neq m$, on calcule *c* comme suit :

$$c = c + \sum_{k=m}^{j^*} N_{i+1}(k-1) \quad (7)$$

Ainsi, pour le cas du LSF_2 on aura:

$$c = 437739667 + \sum_{k=6}^{12} N_3(k-1), \quad \text{soit} \quad c = 483874423.$$

Les mêmes opérations seront réitérées pour les autres coefficients. A ce niveau de la quatrième étape, on transmettra seulement l'indice *c* obtenu à la fin du traitement de la table de LSF_{10} c'est-à-dire *c=483874423* , soit 30 bits, car *c* vaut en binaire:

011100111111001001000101001100.

5. ALGORITHME DE DECODAGE

Au niveau de la réception, le nombre *c* sera réceptionné pour être utilisé dans l'opération inverse, selon l'algorithme de décodage suivant:

Algorithme:

Etape 1. *Trouver* $1 \le j \le 2^{b2}$ *tel que*

$$\sum_{k=1}^{j-1} N_2(L\hat{S}F_{1,k-1}) \le c < \sum_{k=1}^{j} N_2(L\hat{S}F_{1,k-1})$$

Mettre $j^*=j\text{-}1$ *et* $L\hat{S}F_1 = L\hat{S}F_{1,j^*}$

Etape 2. *Calculer* $c = c - \sum_{k=1}^{j^*} N_2(L\hat{S}F_{i,j^*})$

Etape 3. *Pour i=2 à p*

$m = arg\ min_{k=1..2^{bi}} (L\hat{S}F_{i-1,j^*} < L\hat{S}F_{i-1,j^*})$

Mettre j = m si c= 0 sinon

trouver $1 \le j \le 2^{bi}$ *tel que*

$$\sum_{k=m}^{j-1} N_{i+1}(L\hat{S}F_{1,k-1}) \le c < \sum_{k=m}^{j} N_{i+1}(L\hat{S}F_{1,k-1})$$

Mettre $j^* = j\text{-}1$ *et* $L\hat{S}F_i = L\hat{S}F_{i,j^*}$

Si $j^* \neq m$ *calculer* $c = c - \sum_{k=1}^{j^*} N_{i+1}(L\hat{S}F_{i,j^*})$

Etape 4. *Le vecteur LSF reconstitué est donné par* $\{L\hat{S}F_1, L\hat{S}F_2, L\hat{S}F_3, \cdots, L\hat{S}F_p\}$

Dans le cas de l'exemple précédent, *c* a été trouvé égal à *486 314 316.* Dans la première étape de l'algorithme, on cherchera l'indice *j* tel que *c* vérifie :

$$\sum_{k=1}^{j-1} N_2(k-1) \le c < \sum_{k=1}^{j} N_2(k-1) \quad (8)$$

D'après le tableau 1, le représentant du premier coefficient a un indice égal à *6.* D'où $L\hat{S}F_1 = 0.0425$ et, selon l'algorithme, l'indice j^* sera égal à *6 – 1 =5.*

A la deuxième étape on retranchera de *c* la valeur $\sum_{k=1}^{5} N_2(k-1) = 437739667$ pour avoir une nouvelle valeur de *c = 48574649.*

A l'étape trois, une boucle qui commencera à partir du coefficient LSF_2 jusqu'au dernier coefficient LSF_{10} itérera les opérations indiquées dans l'algorithme.

Pour le coefficient LSF_2, par exemple, on effectuera les opérations suivantes. D'abord, trouver *m* tel que $m = M_2(L\hat{S}F_{1,6})$. D'après la table C_2 de la QS du FS1016 [5], *m* vaut *6*. Comme $c \neq 0$, on cherchera *j* vérifiant :

$$\sum_{k=m}^{j-1} N_3(k-1) \le c < \sum_{k=m}^{j} N_3(k-1) \quad (9)$$

j vaut donc 13 selon le tableau 1, car les deux bornes inférieure et supérieure sont égales respectivement à *46134756* et à *4936239*. On met $j^*=j-1=12$ et on attribue à LSF_2 la valeur 0.0838 [5].
De nouveau, $j^* \neq m$, *c* sera alors comparé à *46134756* et ce pour avoir une nouvelle valeur de *c* égale à *2439893 (c-46134756 = 2439893).*

Les mêmes étapes que précédemment seront réitérées pour le reste des coefficients *LSF*.

6. RESULTATS ET INTERPRETATIONS

- *Complexité*

Dans notre cas, nous ajoutons (6x8+4x16)=112 additions et autant de comparaisons pour la partie codage. Ceci correspond 0.007 Mflops, qui reste très faible devant la complexité initiale (environ 8.8 Mflops). Pour le décodeur, 278 additions et 674 comparaisons sont ajoutées, comparativement au décodeur initial, soit 0.032 Mflops.

En ce qui concerne l'occupation mémoire, celle des données augmente légèrement. 110 mots de 32 bits sont nécessaires pour stocker les valeurs entières du tableau 1. Ceci est valable aussi bien pour le codeur que pour le décodeur. Pour la mémoire programme, celle-ci augmente de 10% pour le codeur et de 23% pour le décodeur.

- *Distorsion spectrale*

Des tests sur le standard FS1016, aussi bien avec la QS qu'avec l'application du système d'énumération en treillis, ont été effectués sur deux bases de données de parole, l'une extraite de TIMIT [9] et l'autre Arabe (PAPE) [10]. Pour les deux bases de données, les trames silencieuses des débuts et des fins de phrases ont été limitées à un maximum de 2 trames.
Un total de 218 334 trames de 30 ms a été utilisé. Les performances du quantificateur sont évaluées par la distorsion spectrale moyenne SD qui est souvent utilisée comme mesure objective de la performance d'encodage. SD fournit une bonne corrélation avec la perception auditive humaine.

Le traitement de 218 334 trames a donné une distorsion spectrale moyenne de 1.73 dB, 13.88 pour les % 2-4 dB et 0.22 pour les % supérieurs à 4 dB, confirmant les résultats donnés dans la littérature concernant ce standard [2]-[5].

La sensibilité aux erreurs binaires mérite d'être approfondie ultérieurement. En effet, nous avons constaté une légère dégradation des performances de cette modification du codeur (sans protection) dès que le BER (bit error rate) dépasse 10^{-2} (cas d'un BSC). Les mêmes distorsions ne sont observées qu'aux environs de $5x10^{-3}$ dans le cas du FS 1016 classique.

7. CONCLUSION.

Dans cet article, nous avons présenté le système d'énumération en treillis que nous proposons d'appliquer à la quantification des coefficients LSF en vue de diminuer le débit alloué à ces coefficients dans le standard CELP FS1016, sans changer ses tables de quantification. En fait, nous avons noté que seulement 3.25 % des combinaisons totales que l'on peut former avec les éléments de ces tables de quantification, peuvent vérifier la propriété d'ordonnancement des coefficients LSF et conduisent ainsi à un filtre stable. L'application d'un système d'énumération en treillis nous a permis de diminuer le nombre de bits nécessaires au codage des LSF de 34 bits à 30 bits. Les performances de ce système d'énumération ont été évaluées sur 218 334 vecteurs LSF et ont donné une distorsion spectrale de 1.73 dB, avec des outliers de 13.88 % pour 2-4 dB et 0.22 % pour ceux >4dB, résultat analogue à celui du standard FS1016, mais avec un gain en débit de 133 bits/s.

BIBLIOGRAPHIE

[1] K. Paliwal, W. Kleijn. Quantization of LPC parameters, (Chapter 12). In *speech Coding and Synthesis*. Ed. by W. Kleijn and K. Paliwal, Elsevier, Amsterdam,pp. 433-466, 1995.

[2] ICASSP'95, session: Spectral quantization, 10 articles différents. In *IEEE-ICASSP*, pp. 716-755, 1995.

[3] ICASSP'96. Session: Spectral quantization, 10 articles différents. In *IEEE-ICASSP*, Vol. 1, pp. 737-776, 1996

[4] ICASSP'97. Session: Speech coding at low bit rates, 14 articles différents. In *IEEE-ICASSP*, Vol. 2, pp. 1555-1610, 1997

[5] J.P. Campbell Jr., V.C. Welch, T.E. Tremain. An expandable error protected 4800 bps CELP coder. In *IEEE-ICASSP*, pp. 735-738, 1989.

[6] M.W. Marcellin & T.R. Fischer. Trellis coded quantization of memoryless and Gauss-Markov sources. *In IEEE Trans. on Com.,* Vol. 38, pp. 83-93 Jan 1990.

[7] T. Malone, T. R. Fisher. Enumeration and trellis searched coding schemes for speech LSP parameters. In *IEEE Trans. on speech and audio proc.*, vol.1, N°3, pp. 304-314, July 1993.

[8] F. Itakura. Line spectrum representation of linear predictive coefficients of speech signals. In *JASA*, vol. 57, suppl 1, pp. S35 (A), 1975.

[9] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. In *Linguistic Data Consortium*, 1993.

[10] M. Boudraa, B. Boudraa, B. Guérin. Twenty List of Ten Arabic Sentences for Assessment. I*n Acustica associated with Acta-Acustica,* vol.86, pp. 870-882, 2000.

Evaluation de la qualité vocale dans les télécommunications

M. Guéguin[1,2,3], V. Barriac[1], V. Gautier-Turbin[1], R. Le Bouquin-Jeannès[2,3], G. Faucon[2,3]

[1]France Télécom R&D, TECH/SSTP/MOV, 22307 Lannion Cedex, France
[2]INSERM, U642, Laboratoire Traitement du Signal et de l'Image, Rennes, France
[3]Université de Rennes 1, LTSI, Campus de Beaulieu, 35042 Rennes Cedex, France
marie.gueguin@francetelecom.com

ABSTRACT

This paper is a review of the methods for speech quality assessment. Subjective methods involve human subjects testing systems in various network conditions and voting on an opinion scale. The scores obtained for each condition are averaged to get a mean opinion score (MOS). These subjective tests are the only way to assess perceived speech quality, but they are complex, cost- and time-consuming. Consequently objective methods have been introduced to predict the speech quality as perceived by users. Here, objective methods are classified depending on the context they deal with. This review of objective methods shows a lack of model in the conversational context. Then we propose an objective model of the conversational speech quality, built on a combination of objective models of the listening and talking speech qualities and the delay.

1. INTRODUCTION

Les systèmes de télécommunications sont en constante évolution depuis plusieurs années et nous avons assisté à l'émergence de nouveaux types de transmission, tels que les réseaux mobiles (GSM, bientôt UMTS) et les réseaux de type paquet (*Internet Protocol*, IP). Ces nouvelles technologies sont en pleine expansion du fait de la valeur ajoutée qu'elles apportent aux utilisateurs par rapport à la téléphonie classique, telle que par exemple : la mobilité, ou la possibilité de transmettre non seulement la voix, mais aussi des données et du contenu multimédia, ou encore le coût réduit des appels longue distance. Cependant, contrairement à la qualité de la voix transmise sur le réseau téléphonique commuté (RTC) relativement stable et prévisible, la qualité de service (*Quality of Service*, QoS) de ces nouvelles technologies est généralement non garantie. En effet, elles sont non seulement sujettes à la plupart des dégradations rencontrées avec le RTC (écho, délai, distorsion de l'effet local, bruits, etc.), mais encore introduisent de nouvelles dégradations (distorsion de la parole due au codage, délais augmentés par le traitement numérique), dont certaines sont non linéaires (délai variable appelé « gigue » et pertes de paquets dans les réseaux IP, bruits de fond non stationnaires dans les réseaux mobiles).

Afin de satisfaire leurs clients et de leur offrir la meilleure QoS possible, les opérateurs de télécommunications se doivent de contrôler la qualité perçue par les utilisateurs de leurs services, et doivent pour cela évaluer cette qualité. Les méthodes subjectives, faisant appel à des participants humains qui testent un système dans différentes conditions réelles d'utilisation définies par l'expérimentateur, restent la solution la plus fiable pour évaluer la qualité perçue par les utilisateurs. Bien que ces méthodes subjectives soient le seul moyen d'atteindre le jugement des utilisateurs, les opérateurs de télécommunications cherchent à éviter le recours à de telles méthodes, du fait du coût et du temps qu'elles demandent. Ces méthodes sont décrites dans la section 2.

Ainsi, des méthodes objectives plus poussées que les mesures objectives simples telles que le rapport signal-à-bruit (RSB) et l'erreur quadratique moyenne (EQM) ont été développées. Elles sont construites afin d'être corrélées avec les résultats de tests subjectifs et ainsi constituent un moyen de substitution aux méthodes subjectives. Ces méthodes objectives sont présentées dans la section 3.

Enfin, nous présentons dans la section 4 notre modèle objectif d'évaluation de la qualité vocale en contexte de conversation.

2. TESTS SUBJECTIFS

Lors d'un test subjectif, on demande à des participants de tester un système de télécommunications dans différentes conditions et de noter sur une échelle de qualité la qualité vocale de ce système. D'une manière générale, la qualité dépend de la personne qui la juge. Sa perception met en jeu l'expérience passée, les attentes et l'humeur de chacun. La qualité vocale, dans le cadre des systèmes de télécommunications, est elle aussi dépendante de celui qui l'évalue. Ainsi, les notes des participants pour une condition de test donnée sont moyennées pour obtenir la note moyenne d'opinion (*Mean Opinion Score*, MOS), qui permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale. De plus, la perception de la qualité vocale dépend du contexte et de l'environnement dans lesquels est placée la personne qui juge. En effet, si elle est simplement en train d'écouter un message vocal (contexte d'écoute) ou si elle est impliquée dans une conversation avec un interlocuteur (contexte de conversation), les processus d'attention mis en jeu ne sont pas les mêmes et le jugement de la qualité en est impacté. De même, l'environnement (bruit, informations visuelles ou sonores supplémentaires, etc.) influence le jugement de la qualité. Ainsi, les conditions à tester sont définies en fonction de l'objectif visé, le participant étant amené à évoluer dans un ou plusieurs contextes (écoute, locution et conversation).

2.1. Tests d'écoute

Le principe des tests d'écoute consiste à placer les participants en situation d'écoute et à leur diffuser des séquences audios correspondant la plupart du temps à dif-

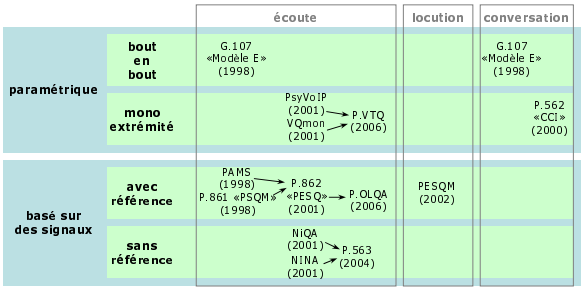


FIG. 1: Les modèles objectifs existants de la qualité vocale

férentes conditions de dégradation. Les conditions testées concernent les dégradations affectant la qualité d'écoute, comme la distorsion de la parole due au codage, le bruit pour l'auditeur et les pertes de paquets. La notation s'effectue selon l'une des méthodes définies par l'Union Internationale des Télécommunications (UIT) dans la Recommandation P.800 [19]. La plus utilisée est la méthode d'évaluation par catégories absolues (*Absolute Category Rating*, ACR) avec les catégories : 5 = Excellente, 4 = Bonne, 3 = Passable, 2 = Médiocre, 1 = Mauvaise. On peut également citer la méthode d'évaluation par catégories de dégradation (*Degradation Category Rating*, DCR) avec les catégories 5 = Dégradation inaudible, 4 = Dégradation audible mais pas gênante, 3 = Dégradation un peu gênante, 2 = Dégradation gênante, 1 = Dégradation très gênante. Plusieurs questions peuvent être posées aux participants permettant ainsi d'évaluer différentes dimensions de la qualité vocale, telles que la qualité globale, le naturel de la voix du locuteur et la dégradation due au bruit.

2.2. Tests de parole et d'écoute

Dans un test de parole et d'écoute, les participants sont placés dans le contexte de locution. Ils doivent donc parler dans le microphone du système à tester et écouter simultanément ce qui arrive du haut-parleur. Cela permet de tester des dégradations affectant la qualité de locution, telles que l'écho, la distorsion de l'effet local et le bruit pour le locuteur. De même que pour les tests d'écoute, les participants notent les conditions testées selon l'une des méthodes définies par les Recommandations P.800 [19] et P.831 [20]. Les questions posées concernent en général la qualité globale, la dégradation due à l'écho et la dégradation due au bruit.

2.3. Tests de conversation

Les tests de conversation sont conçus pour évaluer la qualité dans la situation la plus réaliste. Deux participants sont installés chacun dans une salle et dialoguent via un système de télécommunications. Les conditions testées dans ces tests concernent les dégradations des deux contextes précédents (écoute et locution) ainsi que les dégradations affectant spécifiquement l'interaction de la conversation, comme le délai, la gigue et la double parole. Les conditions testées peuvent être les mêmes pour les deux participants (test symétrique) ou différentes (test asymétrique). Le but étant de reproduire une communication téléphonique réaliste, des prétextes de conversation (sous la forme de dessin à décrire ou de jeu de rôle) sont généralement fournis aux participants. Ainsi, des scénarios de conversation ont été créés [9] ayant pour thèmes par exemple une commande de pizza ou un achat de billet d'avion. Chacun des participants note ensuite la qualité de la conversation qu'il vient d'expérimenter selon l'une des méthodes définies par les Recommandations P.800 [19] et P.831 [20]. Généralement lors de tels tests, il est demandé aux participants d'évaluer la qualité globale, la dégradation due à l'écho, la dégradation due au bruit et l'effort d'interruption. Les tests de conversation sont les plus coûteux et ne permettent pas d'étudier autant de conditions que les tests d'écoute, ils sont donc plus rares.

Quel que soit le type de test subjectif, de nombreuses précautions doivent être prises afin de contrôler les différentes sources de variabilité, telles que le choix des participants, le choix des conditions testées ou l'ordre de présentation des conditions, et afin d'obtenir des résultats fiables et exploitables. Ces tests sont donc fastidieux et coûteux à mettre en place. Les méthodes objectives se présentent comme une alternative aux méthodes subjectives et permettent d'automatiser l'évaluation de la qualité vocale. Cependant, elles doivent avoir une forte corrélation avec les résultats des tests subjectifs, qui représentent le jugement des utilisateurs. Qui dit modélisation objective, dit donc données subjectives pour « alimenter » le modèle.

3. MODÈLES OBJECTIFS

Les modèles objectifs peuvent être classés selon :
– le fait qu'ils se basent sur des mesures physiques du système (paramétriques) ou sur les signaux,
– le besoin qu'ils ont d'avoir accès aux informations des deux côtés du système (bout en bout ou avec référence) ou d'un seul côté seulement (mono-extrémité ou sans référence),
– le contexte dans lequel ils fonctionnent (écoute, locution ou conversation).

La Figure 1 classe les différents modèles existants en fonction de ces trois critères.

3.1. Modèles paramétriques

Les modèles paramétriques utilisent des mesures physiques du système à évaluer pour donner une note de qualité vocale.

Parmi les modèles paramétriques, le modèle E est le plus utilisé. Il a été développé comme un outil bout en bout pour les concepteurs de réseaux et normalisé en 1998 à l'UIT-T dans la Recommandation G.107 [15]. Il a été à la source de nombreux tests subjectifs, qui ont permis de l'optimiser. Le modèle E produit un facteur d'évaluation R compris entre 1 et 100, calculé à partir de mesures physiques des deux côtés du système à évaluer telles que le délai, l'écho, l'atténuation, le bruit de salle, etc. Il peut être utilisé pour estimer la qualité de conversation, et la qualité d'écoute sous réserve de fixer certains de ces paramètres à des valeurs par défaut. Cependant, ce modèle est connu pour donner des résultats faux dans certains cas.

L'équivalent du modèle E en mono-extrémité est le modèle appelé CCI (*Call Clarity Index*), décrit dans la Recommandation P.562 de l'UIT-T [17]. Il permet d'évaluer la qualité de conversation à partir de mesures du système (*e.g.* niveau de parole, niveau de bruit, atténuation de l'écho) effectuées par des sondes non-intrusives appelées

les INMDs (*In-service Non-intrusive Measurement Devices*), décrites dans la Recommandation P.561 de l'UIT-T [16]. Ce modèle permet d'interpréter les mesures faites par les INMDs pour prédire la qualité de conversation, telle que perçue par chaque utilisateur de la communication, en faisant des hypothèses sur le réseau et sur les utilisateurs de chaque extrémité.

Un autre modèle mono-extrémité de la qualité d'écoute, appelé provisoirement P.VTQ [2], est en cours de normalisation à l'UIT-T. Il fixe les objectifs de performances qui doivent être atteintes par des modèles tels que PsyVoIP de Psytechnics [12] et VQmon de Telchemy [5]. Le but de ce type de modèle est de se baser sur les informations des paquets IP sans utiliser les données vocales contenues dans le flot IP (longues à désencapsuler des paquets), afin d'être utilisé dans la surveillance en temps réel de la qualité des réseaux IP. Le modèle estime des paramètres de qualité intermédiaires (taux de perte de paquets, type de perte de paquets et gigue) à partir des informations contenues dans l'en-tête du Real-Time Protocol (RTP). La note de qualité d'écoute est ensuite estimée à partir de ces paramètres.

L'avantage des modèles paramétriques est leur rapidité, ils peuvent donc être facilement embarqués dans des éléments du réseau et les terminaux. Cependant, ils n'atteignent pas les mêmes performances que les modèles basés sur des signaux.

3.2. Modèles basés sur des signaux

Ces modèles, comme leur nom l'indique, utilisent les signaux de référence et dégradé (bout en bout ou avec référence) ou le signal dégradé seul (mono-extrémité ou sans référence) pour prédire la note de qualité vocale du système évalué.

Les modèles avec référence envoient un signal connu (référence) à travers le système à tester, capturent le signal après traversée du système (signal généralement appelé « signal dégradé »), et comparent ces deux signaux afin d'en déduire une note de qualité, qui doit être bien corrélée avec la note MOS.

Parmi les modèles avec référence, les plus utilisés sont les modèles basés sur une comparaison des transformations internes propres à l'oreille humaine, appelés modèles perceptuels. Cette méthode consiste à transformer la représentation physique d'un signal (mesurée en décibels, secondes, hertz) en une représentation psychoacoustique (mesurée en sones, secondes, barks) et est basée sur les travaux de Zwicker et Feldtkeller sur la psychoacoustique [23].

Les modèles perceptuels constituent dorénavant l'approche dominante depuis le développement de méthodes pour évaluer la qualité des signaux audio, qui a abouti à une norme de l'UIT-R [14] : PEAQ (*Perceptual Evaluation of Audio Quality*). Partant de cette norme dans le domaine audio, KPN a développé un outil similaire pour le domaine de la parole appelé PSQM (*Perceptual Speech Quality Measure*) [3], normalisé par l'UIT-T sous le nom P.861 [21]. Cependant, ces deux modèles ont été développés pour évaluer la qualité des codecs audio et vocaux, mais ne sont pas suffisants pour évaluer la qualité d'un système de télécommunications complet, impliquant de nombreuses autres dégradations. En parallèle, British Telecom a développé le modèle perceptuel appelé PAMS (*Perceptual Analysis Measurement System*) [7]. L'avantage de ce dernier est qu'il est plus robuste que PSQM aux délais variables rencontrés en VoIP. Les connaissances de ces deux modèles ont donc été mutualisées et ont permis de créer le modèle PESQ (*Perceptual Evaluation of Speech Quality*), normalisé à l'UIT-T sous le nom P.862 [22]. PESQ permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (perte de paquets, distorsion due au codage et bruit ambiant du côté émission), aboutissant à une corrélation proche de 0,935 avec les données subjectives. Une extension de PESQ au domaine acoustique (avec prise en compte des terminaux) et en bande élargie (de 50 à 7000 Hz, au lieu de 300 à 3400 Hz en bande étroite) est en cours d'étude à l'UIT-T sous le nom provisoire P.OLQA (*Objective Listening Quality Assessment*) [4].

Les méthodes mono-extrémité permettent l'analyse des signaux sans référence connue. L'équivalent de PESQ en mono-extrémité a été normalisé par l'UIT-T sous le nom P.563 [18] à partir des modèles NiQA (*Non-intrusive speech Quality Assessment*) de Psytechnics [13] et NINA (*Non Intrusive Network Assessment*) de SwissQual [8]. Le modèle P.563 permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (distorsion dur aux annuleurs d'écho ou aux systèmes de réduction de bruit, perte de paquets, distorsion due au codage, et bruit ambiant du côté émission), aboutissant à une corrélation proche de 0,89 avec les données subjectives. Le principe de ce modèle est de détecter les trames de parole dans le signal dégradé et d'en extraire un ensemble de paramètres permettant de faire une analyse du conduit vocal et du caractère non naturel de la voix, une analyse des bruits additionnels intenses, une analyse des interruptions, silences et écrêtage temporel. La note de qualité vocale finale est calculée en faisant une combinaison linéaire des différents résultats de l'évaluation de la qualité intermédiaire avec certaines caractéristiques additionnelles du signal.

Tous ces modèles fonctionnent dans le contexte d'écoute. Le modèle perceptuel PESQM (*Perceptual Echo and Sidetone Quality Measure*) [1] évalue la qualité dans le contexte de locution d'un système de communications potentiellement affecté par de l'écho et/ou une distorsion de l'effet local. Ce modèle fonctionne sur le même principe que le modèle PESQ en comparant un signal dégradé avec le signal de référence correspondant. Dans le contexte de locution, le signal de référence est le signal prononcé par le participant dans le microphone et le signal dégradé est le signal retourné par le système dans le haut-parleur du même participant, pouvant donc contenir de l'écho et/ou un effet local distordu. Ce modèle est à la fois un modèle avec référence puisqu'il compare le signal dégradé au signal de référence et un modèle mono-extrémité puisqu'il ne nécessite d'avoir accès aux informations que d'un seul côté du système.

4. MODÈLE OBJECTIF DANS LE CONTEXTE DE CONVERSATION

Comme le montre cet état de l'art des modèles objectifs de la qualité vocale, résumé dans la Figure 1, il n'existe pas encore de modèle non paramétrique de la qualité vocale dans le contexte de conversation. L'UIT-T s'intéresse à la modélisation de la qualité de conversation dans la

Question 20 du Study Group 12, en vue de la normalisation d'un modèle objectif nommé provisoirement P.CQO [10]. L'intérêt existant pour un tel modèle nous a donc conduit naturellement à nous intéresser à sa conception [6]. Lors d'une conversation, chaque utilisateur alterne entre les rôles d'auditeur et de locuteur [11]. La qualité vocale dans ce contexte est donc détériorée par les dégradations rencontrées dans le contexte d'écoute (codage, bruit de fond pour l'auditeur et pertes de paquets) et dans le contexte de locution (écho et distorsion de l'effet local), mais aussi par les dégradations inhérentes à la bidirectionnalité du contexte de conversation, telles que le délai ou la dégradation due au fait que les deux interlocuteurs parlent simultanément (double parole).

Partant de ce constat, notre modèle objectif va donc combiner ces trois composantes de la qualité vocale (écoute, locution et interaction) pour prédire la qualité vocale de conversation correspondante. Pour cela, la relation entre les trois composantes de qualité vocale et la qualité de conversation est déterminée sur le plan subjectif grâce aux notes subjectives correspondantes collectées lors de tests subjectifs. La qualité d'interaction est difficilement évaluable lors de tests subjectifs (il n'existe pas de méthodologie de test normalisée pour ce contexte). Elle est principalement dégradée par le délai. Notre modèle va donc considérer la valeur du délai comme un indicateur de la qualité vocale d'interaction, plutôt que la note subjective d'interaction.

Cette combinaison de trois composantes (qualité d'écoute, qualité de locution et délai) ne consiste pas en une simple juxtaposition (somme, moyenne, etc.). En effet, la qualité en contexte de conversation est plus ou moins influencée par chacune des trois composantes, en fonction des dégradations présentes dans la communication. Ainsi, quand seule une dégradation de la qualité d'écoute est présente, par exemple des pertes de paquets, la note de qualité de conversation sera essentiellement corrélée avec la note de qualité d'écoute, et ne dépendra pas (ou peu) de la note de qualité de locution et du délai. Notre modèle tient donc compte de cette influence du type de dégradation sur la combinaison des trois composantes en introduisant un système de décision, qui pondère l'influence des trois composantes sur la note de qualité de conversation. Ainsi, des tests subjectifs sont nécessaires pour déterminer, en fonction des dégradations, quelle est la relation entre la note de qualité de conversation et les notes de qualités d'écoute et de locution, et le délai.

Une fois déterminée sur le plan subjectif, la relation est transposée sur le plan objectif, en remplaçant les notes subjectives de qualité vocale d'écoute et de locution par des notes objectives fournies par des modèles objectifs, tels que PESQ et PESQM respectivement, et la valeur du délai par sa mesure objective. Le système de décision, déterminé grâce à des tests subjectifs, est alors piloté par les dégradations détectées sur le système de télécommunications évalué.

Appliqué à un test subjectif sur l'écho et le délai, notre modèle aboutit à des corrélations d'environ 0,94 entre les notes objectives et les notes subjectives correspondantes [6].

5. CONCLUSION

Ce papier présente un état de l'art des techniques subjectives et objectives d'évaluation de la qualité vocale. Les méthodes objectives sont classées en fonction de plusieurs critères, notamment celui du contexte dans lequel elles fonctionnent. Ce classement met en évidence le manque de modèles objectifs dans le contexte de conversation, qui est pourtant le contexte le plus courant pour les utilisateurs. Nous présentons donc notre modèle objectif de la qualité de conversation, bâti à partir d'une combinaison des modèles objectifs de la qualité d'écoute et de la qualité de locution et du délai.

RÉFÉRENCES

[1] R. Appel and J. G. Beerends. On the quality of hearing one's own voice. *J Audio Eng Soc*, 50(4) :237–248, 2002.

[2] V. Barriac. Recent standardisation work on non-intrusive evaluation of voice quality in IP environments. In *Proc. CFA/DAGA'04*, pages 53–54, 2004.

[3] J. G. Beerends and J. A. Stemerdink. A perceptual speech-quality measure based on a psychoacoustic sound representation. *J Audio Eng Soc*, 42(3) :115–123, 1994.

[4] J. Berger. Requirements for a new model for objective speech quality assessment P.OLQA. UIT-T COM 12-D.75, 2005.

[5] A. D. Clark. Modeling the effects of burst packet loss and recency on subjective voice quality. In *Proc. IPTEL'01 Workshop*, 2001.

[6] M. Guéguin, R. Le Bouquin-Jeannès, G. Faucon, and V. Barriac. Towards an objective model of the conversational speech quality. In *Proc. ICASSP'06*, 2006.

[7] M. Hollier, A. Rimell, and P. Gray. Verification and use of an auditory perceptual model for subjective analysis of telephone systems. UIT-T COM 12-D.035, 1998.

[8] P. Juric. Non-intrusive speech quality measurement. UIT-T COM 12-27, 2001.

[9] S. Möller. Development of scenarios for a short conversation test. UIT-T COM 12-35, 1997.

[10] J. Pomy. Proposed scope for P.CQO. UIT-T TD 27, 2005.

[11] D. L. Richards. *Telecommunication by Speech : The Transmission Performance of Telephone Networks*. Butterworths, London, 1973.

[12] A. Rix, S. Broom, and R. Reynolds. Non-intrusive monitoring of speech quality in voice over IP networks. UIT-T COM 12-D.49, 2001.

[13] A. Rix and P. Gray. NiQA - Non-intrusive speech Quality Assessment. UIT-T COM 12-D.48, 2001.

[14] Recommandation UIT-R BS.1387. Méthode de mesure objective de la qualité du son perçu, 1998.

[15] Recommandation UIT-T G.107. Le modèle E : modèle de calcul utilisé pour la planification de la transmission, 2003.

[16] Recommandation UIT-T P.561. Dispositif de mesure en service et sans intrusion - Mesures pour les services vocaux, 2002.

[17] Recommandation UIT-T P.562. Analyse et interprétation des mesures en service sans intrusion dans les services vocaux, 2004.

[18] Recommandation UIT-T P.563. Méthode mono-extrémité pour l'évaluation objective de la qualité vocale dans les applications de la téléphonie à bande étroite, 2004.

[19] Recommandation UIT-T P.800. Méthodes d'évaluation subjective de la qualité de transmission, 1996.

[20] Recommandation UIT-T P.831. Evaluation subjective de la qualité de fonctionnement des annuleurs d'écho de réseau, 1998.

[21] Recommandation UIT-T P.861 (supprimée). Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique (300-3400 Hz), 1998.

[22] Recommandation UIT-T P.862. Evaluation de la qualité vocale perçue : méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite, 2001.

[23] E. Zwicker and R. Feldtkeller. *Psychoacoustique : l'oreille récepteur d'information*. Masson, Paris, France, 1981.

La répétition stylistique en anglais oral

Gaëlle Ferré

Université de la Sorbonne Nouvelle - Paris III
EA 1483 "Recherches sur le français contemporain" — Centre de Linguistique Française
13, rue de Santeuil — 75005 Paris, France
Mél: gaelleferre@yahoo.fr

ABSTRACT

This paper is based on a video recording of a face to face interaction between two British girls. In another study on the characteristics of young people's speech involving the same corpus I noticed that one of the specificities of my two speakers lied in the constant repetitions of segments. Some segments are not only repeated in the case of hesitation but also as a stylistic device. I propose to describe in the present paper the stylistic repetitions in terms of what kind of segments are repeated and what is the role of such repetitions. Taking into account lexico-syntactic, prosodic and gestural parameters, I will also show that these repetitions cannot be assimilated to some hesitation on the parts of the speakers.

1. INTRODUCTION

Cet article est parti d'une étude plus large, dans laquelle il s'agissait de décrire les spécificités du discours des jeunes Anglaises (Ferré [9]). L'une des caractéristiques que j'avais alors observée — mais qui ne constitue pas à elle seule une spécificité du "parler jeune" — résidait dans les constantes répétitions de mes locutrices. Or, je fais partie de ces chercheurs qui, à l'instar de Blanche-Benveniste [1], Morel et Danon-Boileau [12], entre autres, travaillent sur l'oral et ses spécificités et accordent une grande importance aux répétitions nombreuses de ce type de corpus, trop souvent considérées comme les "scories de l'oral". Si mon corpus contient un grand nombre de reprises dues à l'hésitation des locutrices, il en contient également d'autres qui n'ont pas pour rôle de faire ressortir le travail de recherche de formulation, bien au contraire.

C'est de ce type de répétition stylistique que je vais parler dans ce papier, répétition qui peut toucher les sons, les mots ou des segments plus larges tels que le groupe intonatif. Après avoir décrit chaque type de répétition et proposé une explication quant au rôle de chacune, je montrerai qu'elles ne peuvent en aucun cas être assimilées aux répétitions dues à l'hésitation. En effet, les indices lexico-syntaxiques, prosodiques, mais aussi mimo-gestuels sont différents dans l'hésitation et dans la répétition stylistique.

Je ne me prononcerai pas en revanche sur le caractère conscient ou inconscient des répétitions.

2. CORPUS

Cette étude repose sur un enregistrement vidéo d'une conversation entre deux jeunes Anglaises, réalisé lors de mon travail de thèse (Ferré [8]). Les deux jeunes femmes de 23 ans ont été enregistrées en studio hors de la présence de tout expérimentateur et avec pour seule consigne de parler comme elles le faisaient régulièrement entre elles. Elles ont été filmées par deux caméras distinctes pendant une demi-heure, puis le corpus a été numérisé et les deux images vidéo montées sur un fichier unique. J'ai ensuite procédé à la transcription et au calcul des paramètres intonatifs à l'aide du logiciel Praat. Que le corpus soit d'une relativement petite taille s'explique par le fait que la transcription précise des gestes des deux interlocutrices doit se faire de manière manuelle et nécessite un temps considérable. Cela n'empêche cependant pas d'y trouver des régularités qui resteront à vérifier sur un corpus plus étendu. Faute de place, je ne présenterai dans cet article qu'un nombre limité d'exemples pour chaque type de répétition, mais l'analyse a été construite sur un nombre d'exemples plus large.

3. LA RÉPÉTITION STYLISTIQUE

3.1. Les répétitions de sons

La répétition de sons — allitération et plus rarement assonance — peut se rencontrer soit au sein d'un seul groupe intonatif, soit à travers deux voire trois groupes intonatifs consécutifs. Ce type de répétition est facilité par la structure phonétique de la langue anglaise mais n'est pas néanmoins sans effet sur le discours. En voici quelques exemples :

(1) it's all (h) {0,38} FU**d**d**y** d**u**d**d**y gran**n**y st**u**ff

(2) it was as if it was gonna (h) **l**eap off the {0,35} WA**LL** // and {0,48} **l**ike {0,16} **L**ATCH itse**l**f onto my **l**eg

(3) GO**T** **t**o **g**o // GO**T** **t**o **g**o **t**onigh**t** // {0,443} GO**T** **t**o **ge**t my **t**icke**t**

où les groupes intonatifs sont séparés par des barres obliques et où les sons répétés apparaissent en gras. Les syllabes toniques (*cf.* Cruttenden [3]) sont notées en petites majuscules et les pauses silencieuses entre

accolades. Ces trois exemples ne sont pas uniques et montrent pour certains une assez grande complexité, comme l'exemple (3), dans lequel sont répétés les sons [g] et [t], mais aussi le premier groupe intonatif dans son intégralité d'abord, puis partiellement dans un deuxième temps. Ce qu'il est intéressant de noter ici, c'est que ces répétitions de sons apparaissent dans des énoncés qui ne sont pas "neutres" sur le plan de la focalisation, mais au contraire dans des énoncés marqués. Elles contribuent donc, par un effet de rythme, à rendre emphatique l'intégralité de l'énoncé, sans qu'il y ait pour autant de focalisation large.

3.2. Les répétitions de mots

La répétition de mots peut se faire de manière consécutive ou non comme le montrent les exemples suivants :

(4) it was *really really* banal

(5) *no no* she said it's an American

(6) and like one week-end *nobody nobody* could have me

(7) oh no there's definitely a *cafe cafe* but there's no {0,274} pub that I've seen

(8) dad's a *Londoner* {0,33} he's a *Londoner* {0,435} born and bred there

Tous les exemples de répétitions de mots ont le même rôle, un rôle d'intensification sémantique. Mais contrairement à l'intensification emphatique de la répétition de sons, on se situe ici dans le domaine de la scalarité. Il s'agit pour les locutrices de formuler un degré plus élevé sur une échelle scalaire (*cf.* Ducrot [4]).

Ce rôle n'est d'ailleurs pas étonnant puisque l'intensification sémantique est l'un des effets les plus fréquents du redoublement ou de la réduplication dans les langues (*cf.* Faits de Langues 29 [6]). L'intensification peut concerner l'aspect qualitatif comme c'est le cas en (4) ou quantitatif du terme intensifé. Ainsi, en (7), c'est l'aspect qualitatif du substantif *cafe* qui est intensifié : il s'agit d'un endroit qui est particulièrement représentatif des cafés dans la classe plus large des débits de boisson. En (6), par contre, c'est la quantité qui est intensifiée, le redoublement de *nobody* étant équivalent à "vraiment personne, pas un chat".

3.3. Les répétitions de groupes intonatifs

Les répétitions de groupes intonatifs peuvent être partielles ou totales, consécutives ou non, avec ajout d'un adverbe ou non. Je ne présenterai ici que des répétitions totales sans modification du groupe intonatif.

(9) I'm gonna stay in London // I don't know who with yet // but (h) I'm gonna stay in London

(10) I heard that Marks and Spencer's are doing really badly they've bought out a whole new (h) fashion range — they are {0,347} // they are

(11) there's this woman there called Madame Maryvonne // and she used to ring round going (h) // who's gonna have Zoe // who's gonna have Zoe // and like (h) // can you have Zoe // can you have Zoe

(12) and they were like // yeah // yeah // we're really enjoying the lessons // we're really enjoying the lessons

Les deux premiers exemples ci-dessus fonctionnent exactement comme les répétitions de mots. Dans l'exemple (9), le 3ème groupe intonatif peut être paraphrasé par "mais une chose est certaine, c'est que je vais rester à Londres" qui vient s'opposer à "I don't know who with yet". Il y a donc un renforcement sémantique par rapport à la première émission. On retrouve ce renforcement dans l'exemple (10) où la répétition montre une plus grande acceptation de la part de l'interlocutrice de l'assertion initiale émise par la locutrice. Cet exemple reste néanmoins ambigu si l'on ne considère que le plan discursif. Seules l'intonation et la mimo-gestualité permettent de l'analyser comme une répétition stylistique et non comme une répétition d'hésitation, comme je le montrerai plus loin.

Dans l'exemple (11), où deux groupes intonatifs sont répétés, on retrouve aussi cette valeur d'intensification : par la répétition, la locutrice veut montrer comment cette femme téléphonait à tout le monde pour que quelqu'un l'accueille dans sa famille le week-end. La répétition crée cependant un autre effet, présent également dans l'exemple (12) et qui consiste en une distanciation ironique de la locutrice entraînée par l'insistance sur l'énoncé. Ce type de répétition apparaît fréquemment dans les passages humoristiques de la conversation (et notamment dans la chute des anecdotes, mais pas uniquement). Par contre, la valeur d'intensification sémantique est toujours présente dans les exemples, quel que soit leur contexte d'occurrence.

4. RÉPÉTITION STYLISTIQUE OU HÉSITATION ?

Reprenons un exemple déjà cité plus haut :

(13) and there this *dead* massive *dead* wasp there

Il pourrait très bien avoir une autre lecture que celle que j'en ai proposé : on peut aussi considérer que la reprise de "dead" est due à un faux départ de la locutrice. En effet, en anglais, la suite "massive dead wasp" est correcte alors que "dead massive wasp" ne

l'est pas, car il y a une contrainte sur la place des adjectifs, la taille devant être exprimée avant d'autres types d'adjectifs ("a large blue shirt/*a blue large shirt"). Pour revenir à notre exemple, on peut donc supposer que la locutrice aurait oublié l'adjectif "massive" qu'elle ajouterait après coup, mais serait alors obligée de reprendre "dead" pour ne pas avoir la suite incorrecte "*dead massive wasp". Une interprétation différente, mais relevant également du concept plus large de l'hésitation (*cf.* Candea [2] ; Duez [5]), pourrait être donnée de l'exemple (6) repris en (14) ci-dessous :

(14) and like one week-end *nobody nobody* could have me

Ici, il n'y aurait pas d'auto-correction, mais la reprise pourrait être due à une hésitation sur ce qui va suivre "nobody".

Pourtant, dans ces deux exemples, comme dans les autres, il ne fait aucun doute, si l'on prend en compte les paramètres lexico-syntaxiques, prosodiques et mimo-gestuels, qu'il ne peut s'agir d'une répétition due à l'hésitation et c'est ce que je voudrais montrer maintenant.

4.1. Les indices lexico-syntaxiques

Afin de distinguer la reprise stylistique de la répétition en contexte d'hésitation, on peut considérer deux indices lexico-syntaxiques : le type de mot repris et la présence ou l'absence d'autres marques du travail de formulation.

La thèse de M. Candea [2] est très claire en ce qui concerne les caractéristiques de l'hésitation. Tout d'abord, les mots repris dans ce contexte sont en général des mots grammaticaux (articles, prépositions, etc.). De plus, il est rare que l'intégralité d'un groupe intonatif, surtout s'il est long, soit repris en cas d'hésitation. Dans mes exemples, on ne peut donc pas compter les répétitions de sons comme des hésitations, ni la plupart des reprises portant sur l'intégralité du groupe intonatif. Quant aux reprises de mots, les répétitions stylistiques impliquent plutôt la reprise d'adverbes, adjectifs, substantifs, soit des mots lexicaux. Il n'en reste pas moins qu'une reprise telle que "they are they are" pourrait, selon ces critères, être comprise comme une reprise d'hésitation, dans un énoncé qui resterait inachevé.

La deuxième caractéristique de l'hésitation qui est mentionnée dans la thèse de Candea est qu'une marque du travail de formulation apparaît rarement seule. En effet, l'hésitation est marquée par différents procédés dont les reprises, les pauses silencieuses et les pauses remplies (*cf.* Swerts [13]), ainsi que les allongements syllabiques. Ceci se vérifie tout à fait dans mon corpus. Or, dans les contextes de répétition stylistique, on ne trouve pas de marques du travail de formulation, notamment aucune pause remplie ni aucun allongement syllabique notoire. Il n'y a pas non plus de pause silencieuse d'hésitation au cœur du groupe intonatif.

4.2. Les indices prosodiques

L'absence de marques du travail de formulation dans le groupe intonatif qui contient une reprise a un effet sur le débit de parole. Alors que le débit de parole est fortement ralenti dans le cas de l'hésitation, il ne l'est pas du tout dans les groupes intonatifs qui contiennent une reprise stylistique par rapport aux groupes environnants. Le débit moyen est aussi plus élevé lorsque le groupe intonatif contient une reprise stylistique que lorsque celui-ci contient une ou des marques d'hésitation (*cf.* Table 1).

Table 1 : Débit moyen (Nb de syll/mn) sur les groupes intonatifs qui contiennent une/des marques d'hésitation comparés aux groupes qui contiennent une reprise stylistique (répétition de mots).

	Hésitation	***Style***
Loc 1	221	386
Loc 2	264	330

En ce qui concerne les pauses silencieuses de focalisation (*cf.* Ferré [7]), on en trouve dans le cas de la répétition de sons (voir les exemples 1-3), où l'allitération/l'assonance vient justement renforcer la focalisation. En revanche, on ne trouve pas de pauses de focalisation dans les deux autres types de répétition. Lorsque la répétition a lieu à travers plusieurs groupes intonatifs, ceux-ci peuvent être séparés en revanche par une pause silencieuse de démarcation.

En ce qui concerne l'intonation des répétitions de mots et de groupes intonatifs, il apparaît que l'intonation du mot ou du groupe répété possède un schéma identique à celle du premier item, tout en respectant la ligne de déclinaison de l'énoncé ou du paragraphe oral (*cf.* Morel et Danon-Boileau [12]).

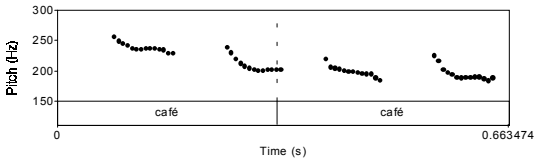


Figure 1 : Courbe intonative (Hz) de la répétition "cafe cafe" tirée de l'exemple (7).

4.3. Les indices mimo-gestuels

Certains paramètres mimo-gestuels peuvent également permettre de différencier une répétition stylistique d'une répétition d'hésitation : je pense notamment à l'orientation de la tête et du regard. Parfois aussi, mais plus rarement dans la simple recherche lexicale, l'hésitation est marquée par un froncement de sourcils par les locutrices (*cf.* Ferré [8]), froncement de sourcils absent en contexte de répétition stylistique.

Dans la conversation, il a été montré que le locuteur ne regarde pas son interlocuteur, alors que celui qui écoute regarde l'autre (*cf*. Kendon, cité par Goodwin [10]). Il y a toutefois des exceptions à cette règle générale, comme par exemple dans le cas de l'interrogation, de la focalisation, où le regard est maintenu sur l'interlocutrice pendant la production de l'énoncé ou de l'item focalisé (Ferré [8]). En revanche, la règle est d'autant plus respectée dans le cas de l'hésitation que la locutrice évite le regard de l'autre. Dans la répétition stylistique (de mots et de groupes intonatifs), la locutrice regarde le plus souvent son interlocutrice sur tout le groupe qui contient la répétition. De plus, l'élément répété même est souvent accompagné d'un geste intensif de la tête ou de la main (geste d'affirmation, de négation, battement de la main — *beat* d'après la terminologie de McNeill [11]), comme le montre cette transcription des gestes de l'exemple (7) :

	there's definitely a cafe cafe
Regard	[regarde l'interlocutrice ___
Tête	⇐ _____[+]
	but there's no {0,274} pub that I've seen
Regard	_____
Tête	⇐ _____

Dans cet exemple, la locutrice regarde son interlocutrice pendant tout l'énoncé, sa tête est tournée vers la gauche également sur tout l'énoncé, mais elle fait un mouvement affirmatif sur "cafe cafe".

5. CONCLUSION

Dans cet article, j'ai montré qu'il existait à l'oral des répétitions stylistiques, telles que les répétitions de sons — assonances et allitérations — dont le rôle est de rendre emphatique un énoncé, mais aussi des répétitions de mots, et des répétitions de groupes intonatifs. Ces deux derniers types de répétition servent principalement à exprimer une intensification sémantique, mais la répétition de groupe intonatif peut également conduire à une prise de distance de la locutrice par rapport à son dire, notamment lorsqu'elle se produit dans des passages humoristiques.

La répétition stylistique se distingue de la répétition en contexte d'hésitation par ses caractéristiques lexico-syntaxiques, prosodiques et mimo-gestuelles, les paramètres les plus fiables de distinction étant :

- absence de marques du travail de formulation autres que la répétition dans le cadre de la reprise stylistique (qui n'est alors pas une marque de TdF),

- débit plus rapide dans la reprise stylistique que dans la reprise d'hésitation,

- regard dirigé vers l'interlocutrice dans la reprise stylistique et association de mouvements de tête/gestes de la main intensifs à la reprise.

Il nous faudra bien évidemment vérifier ces premiers résultats sur un corpus plus étendu.

BIBLIOGRAPHIE

[1] C. Blanche-Benveniste. *Approches de la langue parlée en français*. Ophrys, Paris, 1997.

[2] M. Candea. Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Thèse de doctorat, Paris III - Sorbonne Nouvelle, Directeur : M.-A. Morel, 2000.

[3] A. Cruttenden. *Intonation*. Cambridge University Press, Cambridge, UK, 1997 (Second Edition).

[4] O. Ducrot. *Les échelles argumentatives*. Les Éditions de Minuit, Paris, 1980.

[5] D. Duez. Perception of hesitations in spontaneous French speech. In *Proceedings of The International Congress of Phonetic Sciences*, volume 2, 498-501, Stockholm, 1995.

[6] *Faits de Langues* 29, La réduplication, A. Morgenstern et A. Michaud (coord.), Ophrys, Paris, à paraître.

[7] G. Ferré. Les pauses intra-constituants en anglais spontané, in B. Bel & I. Marlien (eds), *Actes des XXVes Journées d'Étude sur la parole*, 217-220, Fès (Maroc), AFCP, 2004.

[8] G. Ferré. Relations entre discours, intonation et gestualité en anglais britannique. Thèse de doctorat, Paris III - Sorbonne Nouvelle, Directeur : M.-A. Morel, 2004.

[9] G. Ferré. Caractéristiques du parler de jeunes Anglaises. In *Actes du séminaire de recherche "La scalarité : autant de moyens d'expression, autant d'effets de sens*, Bruxelles, 24-25 février 2006, soumis.

[10] C. Goodwin. Conversational Organization. Interaction between Speakers and Hearers. Academic Press, New York, 1981.

[11] D. McNeill. *Hand and Mind : What Gestures Reveal about Thought*. The University of Chicago Press, Chicago and London, 1992.

[12] M.-A. Morel et L. Danon-Boileau. *Grammaire de l'intonation : L'exemple du français*. Ophrys, Paris, 1998.

[13] M. Swerts. Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, 30 :485-496, 1998.

Cohésion temporelle dans les groupes C_1/l/ initiaux en français

Barbara Kühnert[1] et Phil Hoole[2]

[1] Institut du Monde Anglophone & Laboratoire de Phonétique et Phonologie (UMR 7018 – CNRS / Paris 3)
5 rue Ecole de Médecine, 75006 Paris, France
[2] Institut für Phonetik und Sprachliche Kommunikation / Ludwig Maximilians Universität München
Schellingstrasse 3/II, 80799 München, Allemagne
barbara.kuhnert@univ-paris3.fr hoole@phonetik.uni-muenchen.de

ABSTRACT

This work examines aspects of inter-consonantal cohesion within French word-initial C_1/l/-clusters in light of recent proposals of gestural coordination. Based on articulatory and acoustic events, the timing of tongue and lip movements in one subject was studied using an electromagnetic transduction device. More temporal overlap between C_1 and /l/ gesture onset as well as /l/ closure period was found for /pl/ than /kl/. Although matching similar patterns of gestural overlap in initial stop-stop clusters, this 'place of articulation' effect is attributed to low-level motor factors rather than to considerations of perceptual recoverability. An additional analysis of the overall C-centre of /pl/ and /kl/ showed a high temporal stability, confirming a relative constant phasing between initial consonant sequence and following vowel.

1. INTRODUCTION

Dans les modèles dynamiques de la production de parole, les unités pertinentes d'articulation sont des *gestes*, qui sont supposés correspondre aux propriétés phonologiques aussi bien que physiques de la parole [1,2]. Une notion centrale est la *co-production gestuelle* : les gestes articulatoires se chevaucheraient, dans le temps et l'espace, pour produire l'enchaînement de trajectoires qui caractérise la langue parlée.

Une attention particulière a été prêtée, au cours des dernières années, à l'orchestration temporelle des gestes articulatoires. En résumé, il a été montré que cette organisation est influencé par des facteurs globaux, tels que le débit de parole et le registre d'élocution, aussi bien que par des facteurs prosodiques [un aperçu général est proposé par 3 et références citées]. En outre, la coordination temporelle entre gestes semble dépendre de facteurs positionnels. Les groupes consonantiques initiaux de mot (ou de syllabe) tendent à connaître un moindre chevauchement inter-gestes que les mêmes groupes en position médiane de mot ou à travers une frontière de mot [4].

1.1. Coordination inter-gestes dans les séquences consonantiques

Dans le cadre des modèles dynamiques de la production de la parole, les variations observées dans la cohésion gestuelle sont attribuées à des différences dans l'association sous-jacente entre unités gestuelles. Plus spécifiquement, Browman et Goldstein [5] jugent qu'à l'intérieur d'une syllabe les gestes sont associés entre eux à des degrés divers, cette variable étant désignée comme *force d'association (bonding strength)*. En particulier, les gestes consonantiques en position initiale de syllabe présenteraient non seulement une mise en phase avec la voyelle suivante mais aussi une relation de phase entre eux. Le compromis entre les exigences contradictoires de ces deux mises en phase aboutirait à un effet de *C-centre* : un point d'ancrage temporel dans la séquence des consonnes, supposé préserver une synchronisation stable par rapport à la voyelle, indépendamment du nombre et de la nature des consonnes qui composent le groupe. En revanche, aucune mise en phase inter-consonantique n'est censée exister en position finale de syllabe, où l'on s'attend donc à une variabilité plus élevée dans la synchronisation inter-gestes (Nam et Saltzman [6] proposent, dans cette perspective, des simulations de ce type de couplage inter-gestes).

Un autre facteur évoqué pour expliquer des différences de synchronisation inter-gestes est la notion de 'récupérabilité' perceptive des gestes dans la séquence. Ainsi, dans les cas où le chevauchement pourrait masquer des corrélats perceptifs importants de l'une des consonnes de la suite, ce chevauchement sera moindre. Cet argument a été avancé pour expliquer (1) le fait que les groupes de consonnes présentent un moindre chevauchement en position initiale qu'ailleurs, et (2) le fait que dans des séquences de deux plosives le degré de chevauchement gestuel est lié au lieu d'articulation. Un chevauchement moindre est observé dans des suites de consonnes postérieure+antérieure (ex. [kt]) comparées à des suites antérieure+postérieure (ex. [pt]) (voir [4] au sujet de données anglaises, [7] au sujet du géorgien).

Dans ce contexte, et dans la lignée de travaux qui portent sur la coordination temporelle dans les séquences [kl] de plusieurs langues [8,9], notre étude-

pilote s'attache à l'organisation temporelle intra- et inter-gestes de groupes C_1/l/ initiaux en français.

2. METHODE

2.1. Locuteur et corpus

Un locuteur français masculin a participé à l'expérience. Les données examinées ici font partie d'un corpus plus vaste dans lequel le sujet a produit des répétitions multiples d'expressions du type "je vois *mot*$_1$ et *mot*$_2$ et *mot*$_3$". *Mot*$_1$ et *mot*$_2$ sont les mots cibles contenant tous les groupes consonantiques initiaux (C, CC et CCC) phonotactiquement légaux en français. Le *Mot*$_3$ est un distracteur. Les voyelles sont aussi souvent que possible [i] et [a]. Nous examinerons ici dix répétitions de [pl] et [kl] prononcées dans les mots *plaque* et *claque*. Afin d'éviter des effets potentiels de position dans la phrase, chaque mot cible est produit cinq fois comme *mot*$_1$ et cinq fois comme *mot*$_2$. Les phrases étaient présentées sur un écran d'ordinateur et le sujet a été invité de les lire à une vitesse confortable, choisie par lui-même.

2.2. Enregistrements

Les mouvements des lèvres, de la mâchoire inférieure et de quatre points sur la langue ont été capturés à l'aide d'un appareil de transduction électromagnétique à trois dimensions (EMMA 3D). Des capteurs supplémentaires ont été utilisés comme références. Le signal acoustique a été enregistré synchroniquement avec les signaux de mouvement. Une description des procédures d'acquisition, de normalisation et de préparation des données est présentée dans Hoole [10].

Nous nous concentrerons ici sur les données de la lèvre inférieure (LL), de bout de la langue (TT) et du capteur le plus postérieur du dos de la langue (TB), associés respectivement à la production de l'occlusive bilabiale [p], de la liquide [l] et de l'occlusive vélaire [k]. Nous utilisons l'étiquette 'bout de la langue' (TT) de façon conventionnelle mais il faut noter que le capteur était placé à peu près à 1 cm du bout de la langue, ce qui correspond plutôt à la lame de la langue. Le capteur le plus postérieur (TB) a été attaché à peu près à 0,5 cm derrière le point d'occlusion de la langue sur le palais pour une vélaire.

2.3. Analyse de données

Les données ont été analysées à l'aide de scripts MATLAB en considérant les maxima de déplacement vertical et les minima de vélocité tangentielle des trajectoires articulatoires. Les points de mesures, illustrés dans la figure 1, sont, pour [p] : le début du geste de fermeture de la lèvre inférieure, et le début et la fin de l'occlusion bilabiale ; pour [l] : le début du geste de fermeture de la pointe de la langue, et le début et la fin de l'occlusion alvéolaire; et pour [k] : le début du geste de fermeture du corps de la langue, et le début et la fin de l'occlusion vélaire.

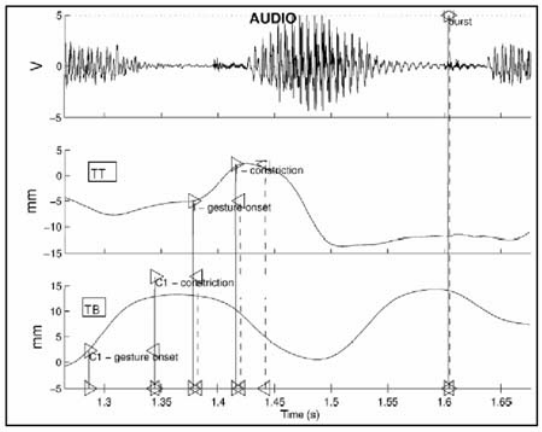


Figure 1: Illustration des mesures dans les signaux EMMA d'une production de *claque* (voir texte pour le détail). Signal acoustique (haut), trajectoire de déplacement vertical de TT (milieu) et de TB (bas).

Deux mesures ont été considérées comme indices de chevauchement temporel entre les gestes des plosives et du [l] : (a) l'intervalle entre la fin de l'occlusion de la plosive C_1 et le début du geste de fermeture de [l] (*onset overlap*), c.-à-d., à quel moment pendant ou après l'occlusion de C_1 est initié le geste du [l]; et (b) l'intervalle entre la fin de l'occlusion de la plosive C_1 et le moment où la constriction de [l] est atteinte (*closure overlap*), c.-à-d., selon la terminologie de Chitoran et al. [7], le délai temporel (*lag*) entre la fin de la constriction de C_1 et le début de la constriction de C_2.

3. RESULTATS

3.1. Chevauchement de gestes et propriétés intra-gestes

Les résultats du chevauchement inter-gestes sont représentés dans la figure 2. Il est à noter que plus la valeur positive est haute, plus le geste du [l] est initié tôt pendant la période d'occlusion de la plosive, donc plus il y a de chevauchement. Inversement, une valeur négative indique que la cible de [l] est atteinte plus tardivement après le relâchement de C_1. Les résultats montrent un effet important du lieu d'articulation de C_1 : le chevauchement entre le début du geste de [l] et la fin de C_1 est bien plus important dans les séquences [pl] que dans les séquences [kl]. De même, la constriction de [l] est atteinte plus tôt après un [p] qu'après un [k]. Un t-test (non apparié, two-tailed), calculé séparément pour les deux indices de chevauchement présente des différences significatives entre les deux contextes [kl] et [pl] (onset overlap : $p > 0{,}001$; closure overlap : $p > 0{,}05$).

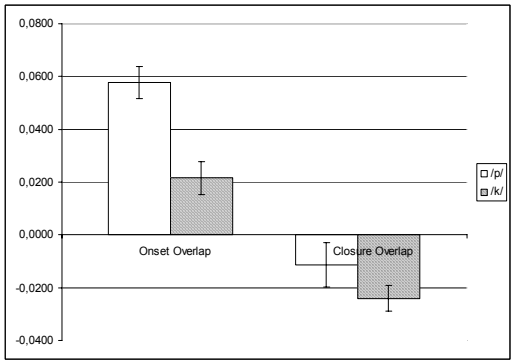


Figure 2: Moyennes et écarts types des intervalles de chevauchement (en ms) pour les deux indices de chevauchement, 'onset overlap' et 'closure overlap' en fonction de la nature de C_1 [p] et [k] (n=10).

Sur la figure 2, on peut également observer implicitement comment les propriétés intra-geste de [l] varient en fonction de C_1, en particulier si on considère l'intervalle temporel entre l'initiation du mouvement de la pointe de la langue et l'atteinte de la cible articulatoire. La table 1 résume les durées des différentes composantes de chaque geste. Si les plosives ont des caractéristiques temporelles assez semblables, les caractéristiques de [l] varient en fonction de la plosive précédente. Dans le contexte [p] la phase de constriction alvéolaire est légèrement plus longue tandis que la durée du geste de fermeture du TT est fortement allongée.

Table 1: Les durées moyennes des gestes de fermeture et des plateaux de constriction maximale en ms (écart type entre parenthèses ; n=10).

	durée du geste de fermeture (ms)	durée du plateau de constriction (ms)
/k/	63 (05)	59 (08)
/p/	67 (08)	55 (12)
/l/, C_1 = /k/	45 (05)	42 (10)
/l/, C_1 = /p/	**69 (06)**	**47 (08)**

3.2. Le timing du C-centre

Comme mentionné plus haut, le C-centre correspond au centre de la séquence consonantique qui est supposé représenter le point dans le temps qui préserve une relation temporelle stable entre la suite des consonnes et la voyelle suivante. Selon Browman et Goldstein [5], le C-centre est calculé comme la moyenne des centres des plateaux de constriction des gestes individuels pour C_1 et C_2. Les C-centres ont été alignés ici par rapport au relâchement acoustique de la consonne finale du mot. Ce point d'alignement est illustré par le curseur sur la droite dans la figure 1. Cette mesure suppose que le geste vocalique, dont les propriétés temporelles sont difficiles à mesurer directement, est coordonné également avec le [k] final dans *claque* et *plaque.* La table 2 montre la localisation moyenne du C-Centre pour les deux groupes consonantiques relativement au point d'alignement. On constate que cette mesure temporelle globale est assez stable entre les deux séquences de consonnes, malgré les différences de coordination temporelle observées ci-dessus entre les gestes consonantique individuels ($p = 0.695$; n.s.)

Table 2: Moyenne et écart type de la localisation du C-centre relatif au burst acoustique du [k] final (en ms; n=10).

Cluster	moyenne (ms)	écart type (ms)
/pl/	268	20
/kl/	264	26

4. DISCUSSION

Les caractéristiques de chevauchement temporel observés pour des groupes C/l/ initiaux en français dans cette étude montrent un effet de lieu d'articulation similaire à celui relaté pour des séquences de plosives initiales [4,7]. En effet, on observe plus de chevauchement dans une suite de consonnes antérieure+postérieure ([pl]) que dans une suite postérieure+antérieure ([kl]). Cependant, il nous semble que ce pattern peut relever non seulement de facteurs de récupérabilité perceptuelle, mais aussi de contraintes simples du système moteur d'exécution.

Premièrement, puisque la production de la liquide [l] n'implique pas une constriction complète dans le tractus vocal, elle ne masque pas les informations perceptuelles éventuelles de la consonne précédente, de la façon dont une plosive le ferait. Ainsi, s'il y en a, les problèmes de récupérabilité perceptuelle sont plus faibles dans le cas d'une production plosive-liquide que dans une production plosive-plosive. De ce fait, plus de chevauchement devrait être permis.

Le deuxième argument repose sur la différence fondamentale observée dans la durée du début du geste jusqu'à l'accomplissement de la cible du [l]. Ce geste de fermeture est significativement plus long lorsque le [l] suit un [p] par rapport à un [k]. Cette différence ne s'explique pas par des différences de spécification gestuelle sous-jacente pour [l], comme par exemple une différence dans le paramètre de raideur qui influencerait le rapport entre la durée, la vitesse et le déplacement d'un geste. Les spécifications des paramètres pour la production d'un [l] sont censées être identiques quelque soit le contexte.

Il nous semble plutôt que le chevauchement limité dans le contexte [k] relève de contraintes sur la configuration globale de la langue pendant la production d'une vélaire. Dans le contexte d'un [p], non-lingual, la langue est libre de se déplacer et

d'anticiper l'articulation du [l] dès la phase de fermeture bilabiale. Par conséquent, le plateau de constriction peut être atteint plus tôt. Les articulations vélaires, en revanche, exigent un mouvement holistique de la langue qui implique également le système bout/lame de la langue. Les caractéristiques spécifiques des plosives vélaires, en particulier, sont connues pour contraindre fortement les mouvements de la totalité de la langue pendant l'intervalle de fermeture, ainsi que l'ont démontré maintes études physiologiques [11,12]. Ainsi, dans le contexte [kl], la lame de la langue ne peut pas exécuter l'articulation du geste de [l] aussi tôt que pour [pl]. Une évidence supplémentaire peut être trouvée dans des productions de [skl] du même locuteur, comme dans *sclérose*. Un exemple représentatif est donné dans la figure 3, recueilli dans la même séance d'enregistrement.

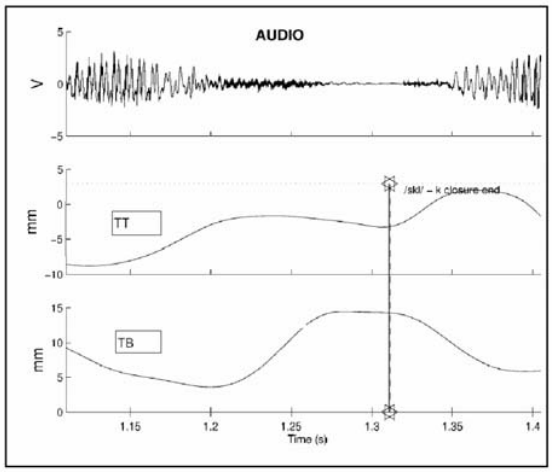


Figure 3: Exemple d'une production de [skl]. Signal acoustique [haut] ; trajectoire verticale de TT (milieu) et de TB (bas). Le curseur est positionné à la fin de la constriction vélaire.

On peut observer que la lame de la langue (au milieu) ne se déplace pas directement de l'articulation de la fricative [s] vers la production latérale pendant l'occlusion vélaire, comme on pourrait s'y attendre. Elle montre plutôt un léger mouvement descendant et le geste du [l] n'est initié qu'à la fin de la constriction de [k].

Bien que certaines des différences temporelles observées n'apparaissent qu'au niveau de l'exécution des articulations, cela ne signifie pas qu'elles ne reflètent pas également des différences dans les spécifications sous-jacentes d'alignement entre les différents gestes. Comme la stabilité du C-centre des groupes de consonnes nous le suggère, il y a un certain réajustement dans la mise en phase des structures consonantiques initiales de sorte qu'une cohésion globale entre séquence consonantique et voyelle soit maintenue.

5. CONCLUSION

Etant donné la variabilité omniprésente dans la production de parole, il va sans dire que nos résultats ne sont que suggestifs pour l'instant. Cependant, ils montrent que seule une investigation détaillée d'une multitude de données empiriques nous permet de déterminer quelles caractéristiques des gestes articulatoires sont les conséquences des propriétés du système moteur et quelles caractéristiques sont dues aux spécifications dans les structures de contrôle central. Nos données montrent également que l'effet du lieu d'articulation sur le chevauchement, i.e. moins de chevauchement temporel dans la direction postérieur+antérieur, reflète dans une certaine mesure les propriétés dynamiques de bas niveau propres aux articulations du dos de la langue.

BIBLIOGRAPHIE

[1] E. Saltzman et K. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333-382, 1989.

[2] C. Browman et L. Goldstein. Articulatory Phonology: an overview. *Phonetica*, 49:155-180, 1992.

[3] D. Byrd. Frontiers and Challenges in Articulatory Phonology. *Proc. ICPhS XV,* pp. 89-92, 2003.

[4] D. Byrd. Influences on articulatory timing in consonant sequences. J*ounal. of Phonetics*, 24(2): 209-244, 1996.

[5] C. Browman et L. Goldstein. Competing constraints on intergestural coordination and self-organization of phonological structures. In *Bulletin de la Communication Parlée*, vol.5, pp. 25-34, 2000.

[6] H. Nam et E. Saltzman. A competitive, coupled oscillator of syllable structure. Proc. *ICPhS XV*, pp. 2253-2256, 2003.

[7] I. Chitoran, L. Goldstein et D. Byrd. Gestural overlap and recoverability: Articulatory evidence from Georgian. In *Papers in Laboratory Phonology VII*, C. Gussenhoven, T. Rietfield and N. Warner (eds.), Mouton de Gruyter, pp. 419-448, 2002.

[8] F. Gibbon, W. Hardcastle et N. Nicolaidis. Temporal and spatial aspects of lingual co-ordination in /kl/ sequences: a cross-linguistic investigation*, Speech and Language*, 36:261-277, 1993.

[9] W. Hardcastle, B. Vaxelaire, F. Gibbon, P. Hoole et N. Nguyen. EMA/EPG study of lingual coarticulation in /kl/ clusters. In *SPM-1996*, pp. 53-56, 1996.

[10] P. Hoole, A. Zierdt et C. Geng. Beyond 2D in articulatory data acquisition and analysis. *Proc. ICPhS XV,* pp. 265-268, 2003.

[11] J. Perkell. Physiology of speech production: results and implications of a quantitative cineradiographic study. MIT Research Monograph 53, MIT Press, Cambridge, MA, 1969.

[12] A. Löfqvist et V. Gracco. Control of oral closure in lingual stop consonant production. *JASA*, 111(6):2811-2827, 2002.

Etude des adductions/abductions totales et partielles des cordes vocales

Chakir Zeroual[1&2], John H. Esling[4] et Lise Crevier-Buchman[2&3]

1. Université Sidi Mohamed Ben-Abdellah, Faculté Polydisciplinaire de Taza, BP. 1223 Taza, Maroc.
2. Laboratoire de Phonétique et Phonologie (UMR 7018) CNRS / Sorbonne –Nouvelle, Paris, France
3. Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75015 Paris, France.
4. Department of Linguistics, University of Victoria, Victoria, BC V8W 3P4 Canada
Chakirzeroual@yahoo.fr ; lise.buchman@numericable.fr ; esling@uvic.ca

ABSTRACT

In this study, we have shown that, in the intervocalic position, simple and geminate voiceless plosives [t tt], present a total abduction of the vocal folds (anterior+posterior parts). While their voiceless emphatic (or uvularized) counterparts [T TT] are produced with an anterior abduction only. Simple and geminate voiced plosives also show a slight anterior abduction which is longer during geminate, and shorter during [gg]. Arguments have been presented showing that the anterior abductions observed during [T TT] and voiced plosives are passive, due to the increase of the intraoral pressure. [i] falsetto shows a slight abduction of the vocal folds that we assign to the action of the intrinsic laryngeal muscles.

1. INTRODUCTION

Les descriptions physiologiques du larynx [1 10 13] montrent que l'adduction totale (parties antérieure+postérieure) des cordes vocales (CV) durant les voyelles et les consonnes voisées est toujours "active" et résulte de la contraction des muscles thyro-aryténoïdien latéral (TAL), interaryténoïdien (IA) et crico-arythéroïdien latéral (CAL). L'abduction totale des CV durant les consonnes sourdes est elle aussi active et résulte de l'action du cricoryténoïdien postérieur (CAP) et de l'absence d'activité du IA. Les analyses physiologiques montrent un autre type d'abduction des CV qui se fait sans la contraction du CAP [8]. Ce genre d'abduction, que nous appelons "abduction antérieure", concerne la partie antérieure de la glotte uniquement et est généralement observé durant la tenue des occlusives voisées [5 11].

L'abduction antérieure durant les occlusives voisées est généralement considérée comme passive et attribuée à l'augmentation de la pression intraorale (Po) [12]. Les modèles aérodynamiques montrent que des ajustements laryngaux et supralaryngaux s'ajoutent durant l'occlusion des occlusives "voisées" pour maintenir une différence d'au moins 2cmH_20 entre la pression sous-glottique (Ps) et Po [1]. Cette condition est nécessaire pour garder le voisement assez longtemps. Ces ajustements se manifestent sous forme (i) d'une augmentation du volume du conduit vocal grâce au relâchement de ses parois qui s'élargissent, au mouvement vers l'avant de la racine de la langue ou à l'abaissement du larynx [14]; (ii) ou d'un relâchement des CV [6] qui facilite leur vibration même avec un faible débit d'air trans-glottique. Sans ces ajustements, Po augmente rapidement d'où non seulement la cessation du voisement, mais aussi l'application d'une force vers le bas sur les CV qui provoque l'abduction passive de leur partie antérieure.

Dans ce travail, nous décrivons, grâce à l'endoscopie, les postures de la glotte durant les occlusives voisées simples et surtout géminées de l'arabe marocain (AM). Nous montrons que les géminées présentent des adductions totales et des abductions antérieures qui ne s'accordent pas avec les prédictions des modèles aérodynamiques. Nous fournirons des données sur l'occlusive coronale emphatique [T] qui possède une abduction "antérieure" même si elle est sourde. Nous montrons également une légère abduction des CV durant le mode "falsetto".

Nous discutons, principalement, si les abductions particulières de la glotte que nous avons en AM sont actives (produites par la contraction d'un muscle intrinsèque du larynx) ou passives (résultant notamment de l'augmentation de Po). Pour cela, nous comparons nos résultats physiologiques et acoustiques avec les données d'autres langues rapportées par les auteurs.

2. MATERIEL ET METHODE

Des enregistrements vidéo ont été effectués grâce à une caméra (Olympus OTV-SF, 25images/secondes) reliée à un endoscope (Olympus Enf-P3) inséré à travers les fosses nasales d'un locuteur marocain (38 ans). Une partie de l'enregistrement (partie (i) du corpus) a été enregistrée en utilisant aussi la stroboscopie.

Ce locuteur a répété sept fois un corpus composé de 3 parties. (i) [i] tenue prononcée avec des qualités de voix différentes (modale, falsetto, chuchotée, etc), (ii) presque toutes les consonnes de l'AM prononcées dans des mots et non mots sous la forme simple ([-iCi]) et géminée ([-ɪCCi-]). Nous présentons ici les résultats des occlusives simples et géminées [b t d T D g] ([T D] : coronales emphatiques (uvularisées) correspondantes de [t d]), ainsi que ceux de la voyelle tenue [i] modale et falsetto .

Nous présentons aussi des observations effectuées durant un enregistrement séparé par transillumination, où le même locuteur a prononcé un corpus très similaire à celui de la présente expérience par endoscopie. Nous donnerons quelques commentaires sur les tracés obtenus par transillumination de [t tt T TT], l'analyse des autres consonnes sera présentée dans un travail séparé.

Les films vidéo ont été analysés par Adobe Premiere7, les images par Adobe Photoshop7, les données audio par Praat, les courbes de l'aperture glottique par Matlab et les analyses statistiques par StatView.

3. RESULTATS ET DISCUSSION

3.1 La voix modale et falsetto (Figure 1)

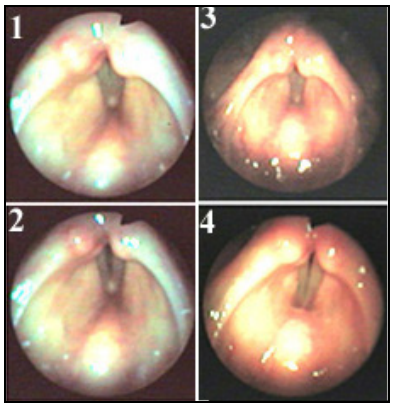


Figure 1 : Postures de la glotte durant : les phases fermées (1) et ouverte (2) de [i] tenue modal (image par stroboscopie), [i] modal (3) et falsetto (4) .

Nos observations à l'aide de la caméra 25images/seconde durant [i] tenue modale (F0=200Hz), montrent que les CV restent fermées sur toute leur longueur. Grâce à la stroboscopie, nous avons pu identifier durant [i] modale une alternance entre des phases fermées et ouvertes de la glotte. Pendant la phase ouverte, aucun écartement au niveau des aryténoïdes n'a été enregistré. Durant [i] tenue produite avec le mode falsetto (F0=350Hz), la glotte apparaît plus allongée, et les CV ne sont pas aussi rapprochées entre elles que durant la voix modale.

Les descriptions physiologiques [1 10 13] montrent que la voix "modale" s'accompagne des contractions modérées des muscles TAL et CAL qui attirent le cartilage aryténoïde vers l'avant et vers l'intérieur du larynx. D'où une compression antérieure-postérieure de la glotte et une fermeture de sa partie antérieure. La contraction modérée du muscle IA permet la fermeture de la partie postérieure de la glotte ; celle du muscle cricothyroïdien (CT) empêche un raccourcissement important des CV et une augmentation excessive de leur masse et ajuste leur tension. Selon la théorie myoélastique [13], la vibration des CV est un processus passif conséquence de la combinaison des forces actives exercées par les muscles adducteurs et par la pression sous-glottique. Le fait que durant la phase ouverte du [i] modale les aryténoïdes sont collés s'accorde avec les descriptions de cette théorie.

Selon Hirano [cité dans 11], chaque CV est constituée de plusieurs couches souvent réduites à deux : la couche externe ou la muqueuse et la couche interne ou le corps (cover et body). Durant la voix modale, VOC (muscle vocal) se contracte et exerce une force antagoniste au CT, la tension du corps augmente et celle de la muqueuse baisse. Par contre, durant falsetto, la contraction du CT est plus marquée, alors que le VOC reste inactif, d'où un allongement excessif des CV, une réduction de leur épaisseur et une augmentation de la tension aussi bien du corps que de la muqueuse. Le résultat est une augmentation excessive de F0 qui caractérise le mode falsetto. Les données rapportées par les auteurs ainsi que les nôtres, montrent que falsetto est souvent accompagnée d'une petite ouverture de la glotte. Cette dernière est attribuée à l'activité très marquée du CT [11], ainsi que probablement à l'absence de contraction du VOC qui garde les CV relativement éloignées.

3.2 Les occlusives sonores (Figure 2)

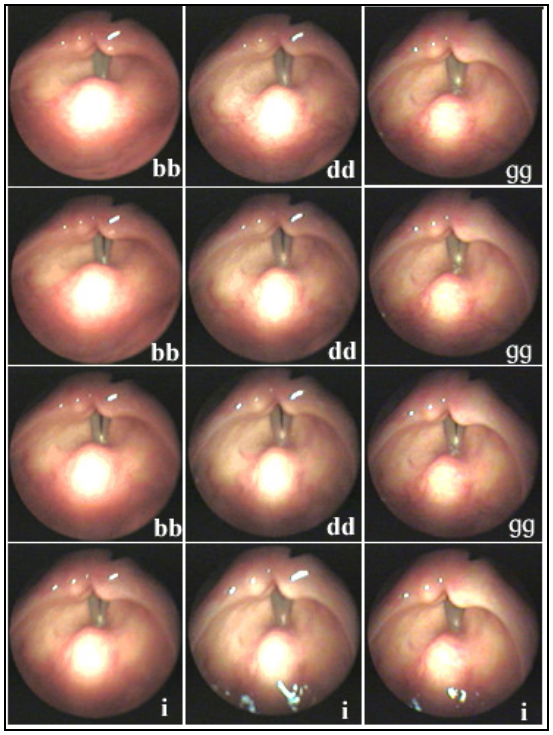


Figure 2 : Postures de la glotte durant [-ɪbbi-], ([-ɪddi-]) et [-ɪggi-] produites dans des mots de l'AM.

[b d D g] ont une adduction des CV pendant toute la durée de l'occlusion. Durant certaines de leurs occurrences, une légère abduction antérieure est observée une image avant la voyelle. Durant [gg] la glotte peut rester fermée durant toute l'occlusion ou s'ouvrir une image seulement avant la voyelle. Par contre, durant [bb dd DD], les CV sont entièrement adduquées durant leur première moitié, une abduction antérieure est observée généralement pendant deux images avant la voyelle suivante (Fig. 2).

Le voisement peut se maintenir durant toute la durée de l'occlusion de [b d D g] ou cesser immédiatement avant le relâchement. Durant les géminées, le voisement se maintient rarement au delà de la première moitié de l'occlusion durant [bb dd DD], et occupe pratiquement toute la durée de l'occlusion durant [gg] (Tab. 1). La

durée du voisement et du silence (tableau 1) sont respectivement les plus longues et les plus brèves durant [gg], et les plus brèves et les plus longues durant [bb]. Notons que la durée de l'occlusion durant les simples est très inférieure comparée aux géminées. L'occlusion de [bb dd DD], apparaît plus longue que celle de [gg].

Des expériences par EMG [7], montrent que, comparée aux voyelles adjacentes, l'occlusive voisée développe à l'intervocalique une baisse assez importante de l'activité de INT et VOC et légère de LCA, alors que CT semble garder le même niveau de contraction. Rappelons que la contraction très importante du CT combinée à l'absence d'activité du VOC, comme durant falsetto, peut induire l'abduction antérieure des CV. Cette dernière, généralement observée avant le relâchement des occlusives voisées, ne peut être due aux actions inverses du CT et VOC, puisque l'activité du CT doit rester faible.

L'abduction antérieure ainsi que le silence durent plus longtemps durant [bb dd DD] que durant [b d D], probablement parce que Po atteint une amplitude importante longtemps avant le relâchement des premières. Le fait que l'abduction et le voisement sont plus longs durant [gg] comparé à [DD] et surtout à [bb DD] suggèrent aussi que Po atteint une valeur proche de Ps beaucoup plus tôt durant [DD] et surtout [dd bb]. Cette déduction ne s'accorde pas avec les descriptions aérodynamiques proposées aux consonnes occlusives qui prédisent une augmentation plus rapide de Po et un voisement plus court durant [gg] et l'inverse durant [bb]. Des résultats inattendus ont été relevés aussi par Cohn et al. [2] dans des langues austronésiennes où le voisement occupe toute la durée de la phase d'occlusion de toutes les occlusives voisées géminées.

Des ajustements articulatoires supplémentaires semblent se développer de manière plus importante durant l'occlusion de [gg], ce qui permet d'avoir un voisement qui peut durer jusqu'à 140 msec. Des hypothèses perceptives peuvent aussi expliquer les différences par rapport à la durée du voisement entre [bb dd DD gg]. L'AM possède [b bb] mais pas [p pp], les propriétés acoustiques de [b bb] peuvent donc être assez variables sans que cela n'affecte leurs catégories phonologiques. Par contre, une durée plus brève du voisement durant [gg] facilite l'augmentation de Po, d'où un burst qui serait plus long et surtout plus intense, donc moins distinct de celui de [kk].

Nos données combinées avec les observations antérieures par EMG plaident en faveur de l'hypothèse qui considère l'abduction antérieure durant les occlusives voisées simples et géminées comme étant passive.

3.3 Les occlusives sourdes [t] et [T] (Figure 3)

Durant [t tt], les aryténoïdes s'écartent progressivement, puis se rapprochent de nouveau tout aussi progressivement (Fig. 3). Durant [T TT], les aryténoïdes restent collés, alors que la partie antérieure de la glotte s'ouvre très légèrement dès le début de l'occlusion pour garder cette posture jusqu'au début de la voyelle suivante. Notons que la durée du VOT est très importante durant [t tt] (65msec) et très faible durant [T TT] (21msec).

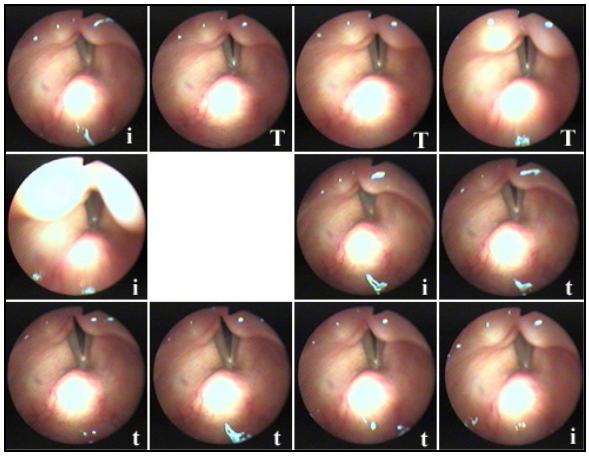


Figure 3: Postures de la glotte durant une partie des items [iTib] et [itih].

Les tracées par transillumination (figure 4) montrent que seules [t tt] possèdent des phases d'abduction et d'adduction qui sont bien représentées. La courbe de l'aperture glottique reste presque plate durant [T TT].

L'abduction totale des CV durant les consonnes sourdes à l'intervocalique est généralement obtenue grâce à la contraction du CAP et la non activation du IA. Hirose et al [7] ont montré que la contraction du CAL durant l'occlusive sourde à l'intervocalique est plus faible que durant la voyelle, mais pratiquement identique à sa correspondante voisée. Dixit [4] montre une contraction plus marquée du CT durant l'occlusive sourde que durant la voisée, mais ce résultat n'a pas été toujours obtenu par d'autres travaux. Il semble que LCA est actif pendant la sourde pour éviter une ouverture excessive de la glotte. Plus généralement, la contraction du CAL permet surtout d'établir la glotte dans le mode de "parole" ou "préphonatoire" [11], alors que les autres muscles du larynx contribuent aux différences segmentales (CAP, IA) ou suprasegmentale (CT, VOC) par rapport aux modes phonatoires ou au registre de la voix.

[T] partage avec l'occlusive sourde non aspirée [t*] du coréen dite fortis la durée très faible du VOT et probablement d'autres propriétés physiologiques. Selon Kagaya [9], durant [t*] à l'intervocalique, la partie postérieure de la glotte reste adduquée, une abduction antérieure est observée durant le début de l'occlusion suivie d'une adduction totale des CV qui s'établit avant le relâchement. Des études par EMG [8] ont montré qu'à l'intervocalique, CAP est inactif durant toute la durée de [C*], alors que VOC se relâche durant le début de l'occlusion et se contracte fortement avant le relâchement. Les modélisations aérodynamiques de Dart [3] suggèrent que [t*] est produite avec une tension plus importante des parois du conduit vocal. Cette dernière combinée à

l'absence d'activité du VOC durant le début de l'occlusion, comme c'est le cas durant [C*] et probablement durant [T], faciliterait l'abduction passive de la glotte induite par l'augmentation rapide de Po. Une activation anticipée du VOC combinée à l'absence de contraction du CAP accélèrent l'adduction totale des CV, d'où la durée très brève du VOT de [t*] et [T].

Tableau 1 : Durées et écartypes en msec (1 locuteur x 7 répétitions) du voisement, silence (absence de voisement) et de l'occlusion durant [bb dd DD gg], de l'occlusion et du VOT de [tt TT] produites dans le contexte [ɪCCi].

	Voisement	Silence	Occlusion	VOT (+)
[bb]	70 (18)	93 (19)	163 (10)	
[tt]			142 (10)	63 (3)
[dd]	83 (16)	77 (9)	160 (9)	
[TT]			175 (13)	15 (2)
[DD]	107 (17)	61 (21)	169 (8)	
[gg]	138 (13)	9 (13)	145 (4)	

4. CONCLUSION

Cette étude montre qu'en position intervocalique, l'occlusive sourde [t], simple et géminée, présente une abduction totale des cordes vocales (parties antérieure+postérieure). Par contre, ses correspondantes sourdes emphatiques [T TT] présentent une abduction de la partie antérieure de la glotte uniquement. Les occlusives voisées, simples et géminées, développent elles aussi une légère abduction antérieure qui dure plus longtemps durant les géminées, elle est aussi plus courte durant [gg]. Des arguments ont été présentés qui montrent que l'adduction antérieure observées durant [T TT] et les occlusives voisées est passive, due à l'augmentation de la pression intraorale. [i] falsetto développe elle aussi une légère abduction des cordes vocales que nous attribuons à l'action des muscles intrinsèques du larynx.

BIBLIOGRAPHIE

[1] J.C. Catford. *Fundamental Problems in Phonetics*. Edinburgh University Press, Edinburgh, 1977.

[2] A.C. Cohn, W.H. Ham, R.J. Podesva. The phonetic realization of singleton-geminate contrasts in three languages of Indonesia. *Proceedings of the XIVth ICPHS*, San Francisco: 587-590, 1999.

[3] S.N Dart. An aerodynamic study of Korean stop consonants: Measurements and modeling. *J. Acoust. Soc. Amer.*, 81(1): 138-147, 1987.

[4] R.P. Dixit. *Neuromuscular aspects of laryngeal control: with special reference to Hindi*. Ph. D. dissertation. University of Texas Austin, 1975.

[5] R.P. Dixit. Glottal gestures in Hindi plosives. *Journal of Phonetics* 17: 213-237, 1989.

[6] M. Halle and K. N. Stevens. A note on laryngeal features. QPR *Quarterly Progress Report, Research Laboratory of Electronics*, MIT 101: 198-213, 1971.

[7] H. Hirose and T. Ushijima (1978) Laryngeal control for voicing distinction in Japanese consonant production. *Phonetica* 35: 1-10, 1978.

[8] K. Hong, S. Niimi and H. Hirose. Laryngeal Adjustments for the Korean Stops, Affricates and Fricatives - An Electromyographic Study. *Ann. Bull. RILP*. 25: 17-31, 1991.

[9] R. Kagaya. Laryngeal gesture in Korean stop consonants. *Ann. Bull. RILP* 5: 15-23, 1971.

[10] J. Laver. *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[11] M. Sawashima, and H. Hirose. Laryngeal gestures in speech production. *Ann. Bull. RILP*. 14: 29-51, 1980.

[12] K.N. Stevens. Vocal-fold vibration for obstruent consonants. In J. Gauffin and B. Hammarberg (eds.) Vocal fold physiology. *Acoustic, perceptual, and physiological aspects of voice mechanisms*. Singular Publishing Group, San Diego: 29-36, 1991.

[13] J.W. Van Den Berg. Mechanism of the larynx and the laryngeal vibrations. in B. Malmberg, (ed.) *Manual of Phonetics*. Amsterdam, North-Holland: 278-307, 1968.

[14] J. R. Westbury. Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *J. Acoust. Soc. Am*. 73: 1322-1336, 1983.

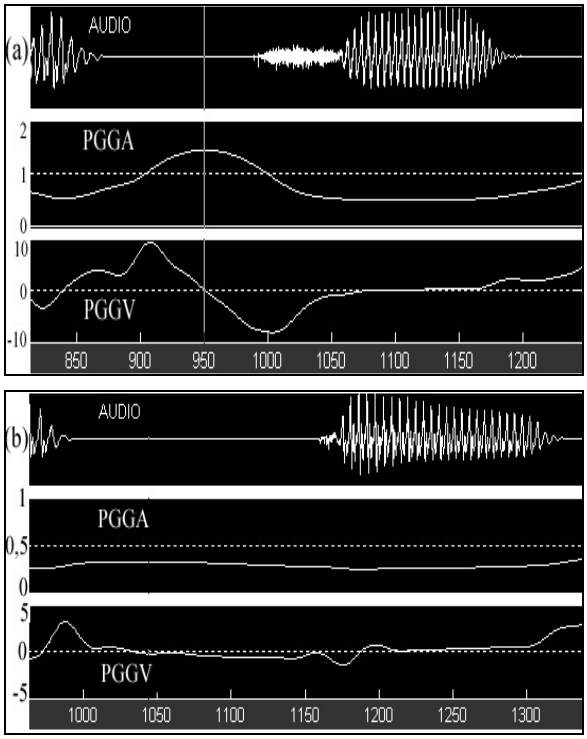


Figure 4 : Tracés de la courbe de l'aperture glottique (PGGA) obtenue par transillumination, ainsi que la courbe de la vélocité (PGGV) calculée sur la première durant les séquences [-ɪtti-] (a) et [-ɪTTi-] (b) extraites des items *[bɪttit] et [bɪTTit]. Notons que l'axe des 'y' de PGGV de (a) et (b) n'a pas la même échelle.

Session XIX

Conférence Invitée

Vendredi 16 juin 2006 - 09h00 10h00

Imagerie cérébrale du bilinguisme et de l'apprentissage des langues

Christophe Pallier

CNRS, INSERM, Unité 562 de Neuroimagerie Cognitive,
Service Hospitalier Frédéric Joliot, CEA, DSV, DRM
4 place du Général Leclerc, F-91401 Paris
christophe.pallier@m4x.org

ABSTRACT

The neurolinguistics of bilinguism and language acquisition is still in infancy. This paper presents a quick overview of some brain imaging studies of language acquisition conducted in our lab. We describe a study showing that, as grammatical skills in L2 increase, the cerebral activations elicited by sentence building in L1 and in L2 became more and more similar. In another series of studies, we discovered anatomical and functional cerebral correlates of the abilities to memorize, perceive or produce foreign speech sounds.

1. CONVERGENCE ENTRE LES REPRÉSENTATIONS CORTICALES DE LA PREMIÈRE ET DE LA SECONDE LANGUE

Les deux langues d'un bilingue sont-elles sous-tendues par les mêmes circuits neuronaux ou bien par des réseaux distincts ?[1]

Une série d'études d'imagerie cérébrale du bilinguisme ont employé la tomographie par émission de position (TEP) ou la résonance magnétique (IRM) pour mettre en évidence les patterns d'activation cérébrale associés au traitement de la première (L1) ou de la seconde langue (L2), lorsque les sujets accomplissent des tâches telles que la lecture de mots, la dénomination d'image, la compréhension de phrases écrites ou orales... La majorité de ces travaux décrivent des activations très similaires pour les deux langues (cf. [9, 10] pour des revues de la littérature), résultat généralement interprété comme montrant que les mêmes circuits sont employés pour traiter l'une où l'autre langue.[2] Néanmoins, certaines études ont décrit des activations partiellement distinctes pour L1 et L2 chez des sujets ayant un niveau intermédiaire dans la seconde langue et/ou ayant apprise celle-ci tardivement (après 10 12 ans) [8, 2].

Ces observations suggèrent que la représentation corticale de L2 (c.-à.-d. les aires recrutées par L2) devient de plus en plus similaire à celle de L1 lorsque l'apprentissage de L2 progresse. Pour tester cette hypothèse, nous avons scanné en IRMf dix sujets français ayant un niveau moyen en anglais et un éventail de scores au sous-test grammatical du TOEFL (Test of English as a Foreign Language) [4]. Dans le scanner, les participants devaient soit lire des listes de mots, soit construire des phrases à partir de ces mêmes listes. La soustraction entre les activations cérébrales associées à ces deux conditions montre l'implication du gyrus frontal inférieur gauche (l'aire de Broca) dans la construction des phrases [7] Pour chaque sujet, nous avons localisé à l'intérieur de cette région les lieux où l'activation cérébrale était maximale, en anglais d'une part, et en francais d'autre part. La distance entre ces deux maxima d'activation nous a servi à estimer la distance entre les représentations de L1 et L2. Conformément à l'hypothèse du rapprochement de L1 et L2, il y avait une corrélation significative entre les scores individuels au test du TOEFL et la distance entre les maxima d'activation en anglais et en francais : les sujets ayant les meilleurs scores avaient les maxima les plus proches. Ce résultat suggère que plus le niveau de maitrise de la seconde langue augmente, du moins en ce qui concernent la grammaire, plus les activations cérébrales associées à la construction de phrases deviennent similaires. Si tel est le cas, une étude longitudinale devrait donc montrer un déplacement des activations de la seconde langue vers celle de première.

2. BASES CÉRÉBRALES DES DIFFÉRENCES INTERINDIVIDUELLES

2.1. Différences fonctionnelles

Il existe une variabilité interindividuelle non négligeable dans la capacité à apprendre une seconde langue. De nombreux facteurs interviennent certainement : langues déjà connues, habilité à imiter, motivation... En particulier, une étude menée chez des enfants apprenant l'anglais a montré que leurs scores dans des tâches de mémoire phonologique (mémorisation et répétition de pseudo-mots) étaient prédictifs de leur performance en anglais deux années plus tard [11].

Inspirés par l'idée que la capacité de représenter et de mémoriser des mots étrangers était déterminante pour apprendre une seconde langue, nous avons comparé deux groupes de personnes ayant grandi dans un environnement bilingue (à Singapour) mais différant dans leur maîtrise de la seconde langue [1]. Nous les avons scannés alors qu'ils écoutaient des séries de mots français (une langue qui leur était inconnue) et devaient détecter lorsqu'un item était répété. Dans cette tâche, qui implique les processus de mémoire phonologique mais qui demeure assez facile, les performances des deux groupes étaient similaires. Toutefois, de manière remarquable, les activations cérébrales

[1] Ici, un bilingue est défini comme une personne pouvant utiliser deux langues, sans préjuger du niveau atteint dans ces deux langues. Autrement dit, nous ne reservons pas le terme "bilingue" aux personnes ayant un niveau quasi natif dans les deux langues.

[2] Il faut garder à l'esprit la résolution actuelle des images d'IRMf, de l'ordre de $3 \times 3 \times 3$ mm^3.

étaient différentes entre les deux groupes : les meilleurs bilingues recrutaient plus intensément les régions de l'insula et du gyrus frontal inférieur gauche, impliqués dans la production de la parole, alors que le groupe de bilingues moins équilibrés avait des activations révélant un effort attentionnel plus important. Une interprétation possible est que les personnes qui ont un meilleur niveau dans leur seconde langue utilisent leur circuits de mémoire phonologique de manière plus efficace. Notre étude ne permet toutefois pas de savoir si cette caractéristique était présente avant l'apprentissage de la seconde langue ou bien si elle en est le résultat. Pour cela, il faudrait pouvoir mener une étude avant et après l'apprentissage.

2.2. Différences anatomiques

L'étude précédente révélait une différence d'utilisation des ressources cognitives en fonction du niveau de bilinguisme. Des caractéristiques anatomiques peuvent-elles également expliquer la plus ou moins grande facilité à percevoir des différences entre des phonèmes d'une langue étrangère ou à les articuler ? Par exemple, le contraste entre les consonnes dentales et retroflexes en Hindi est assez difficile à entendre pour des sujets francais. Narly Golestani, en postdoc dans notre laboratoire, en a testé une soixantaine et a constitué deux groupes en fonction de la rapidité avec laquelle ils avaient appris à distinguer des syllabes utilisant le contraste dental-retroflexe [5]. Ensuite, nous avons mesuré chez chaque sujet les volumes des gyri de Heschl gauche et droit, structures qui abritent le cortex auditif. L'analyse des volumes a révélé que les personnes qui avaient eu le plus de facilité pour distinguer les syllabes hindi avaient en moyenne des cortex auditifs gauches plus volumineux que ceux ayant eu plus de difficultés. La différence était significative pour le volume de matière blanche, reflètant potentiellement un plus grand nombre ou une meilleure myélénisation des fibres du cortex auditif. On peut spéculer que ces paramètres influencent la précision de la représentation temporelle des sons, particulièrement utile pour discriminer des contrastes consonantiques associés à des transitions acoustiques rapides.

Dans une seconde expérience, la capacité des sujets à répéter une consonne étrangère à été évaluée [6]. Nous avons choisi une consonne uvulaire du Farsi, facilement distinguable des phonèmes français. Les sujets devaient la produire dans différents contextes et deux locutrices natives du Farsi leur ont attribué des scores de qualité d'accent. Ces scores ont ensuite été corrélés avec les images individuelles de probabilité de matière blanche ou de matière grise (suivant la technique dite "voxel-based morphometry"). Cette analyse a révélé que plus l'imitation était fidèle, plus le sujet avait une probabilité importante d'avoir de la matière blanche dans deux aires classiquement impliquées dans la mémoire phonologique et dans l'articulation : le cortex pariétal inférieur et l'insula.

3. CONCLUSION

Les expériences que nous avons présentées doivent être considérées comme des premières tentatives d'exploration d'un domaine encore très mal connu : les bases cérébrales de l'acquisition d'une langue étrangère. Il faut espèrer que ces résultats encourageants seront confirmés et généralisés à des populations plus larges de sujets et à divers types de situation. Il est important de souligner que le fait de mettre en évidence des différences entre les cerveaux des personnes qui ont des facilités d'acquisition et celles qui en ont moins ne signifie pas que ces dernières doivent être découragées d'apprendre. En effet, même chez l'adulte, un entraînement peut produire des modifications corticales macroscopiques détectables [3].

4. RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES

[1] Michael Chee, Chun Siong Soon, Hwee Ling Lee, and Christophe Pallier. Left insula activation : A marker for language attainment in bilinguals. *Proceedings of the National Academy of Sciences USA*, 101(42) :15265–15270, 2004.

[2] S. Dehaene, E. Dupoux, J. Mehler, L. Cohen, E. Paulesu, D. Perani, P.-F. van de Moortele, S. Léhericy, and D. Le Bihan. Anatomical variability in the cortical representation of first and second languages. *Neuroreport*, 8 :3809–3815, 1997.

[3] Bogdan Draganski, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Bogdahn, and Arne May. Neuroplasticity : changes in grey matter induced by training. *Nature*, 427(6972) :311–312, Jan 2004.

[4] Narly Golestani, F.-Xavier Alario, Sébastien Meriaux, Denis LeBihan, Stanislas Dehaene, and Christophe Pallier. Syntax production in bilinguals. *Neuropsychologia*, 2006, in press.

[5] Narly Golestani, Nicolas Molko, Stanislas Dehaene, Denis Lebihan, and Christophe Pallier. Brain Structure Predicts the Learning of Foreign Speech Sounds. *Cerebral Cortex*, Apr 2006.

[6] Narly Golestani and Christophe Pallier. Anatomical correlates of foreign speech sound production. *Cerebral Cortex*, 2006, in press.

[7] P. Indefrey, C. M. Brown, F. Hellwig, K. Amunts, H. Herzog, R. J. Seitz, and P. Hagoort. A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences USA*, (98) :10, 5933–5936 2001.

[8] Karl H. S. Kim, Norman R. Relkin, Kyong-Min Lee, and Joy Hirsch. Distinct cortical areas associated with native and second languages. *Nature*, 388(10 july 1997) :171–174, 1997.

[9] Christophe Pallier and Anne-Marie Argenti. Imagerie cérébrale du bilinguisme. In Olivier Etard and Nathalie Tzourio-Mazoyer, editors, *Cerveau et Langage. Traité de Sciences Cognitives*, pages 183–198. Hermès Science, Paris, 2003.

[10] D. Perani and J. Abutalebi. The neural basis of first and second language processing. *Current Opinion in Neurobiology*, 15(2) :202–206, 2005.

[11] E. Service. Phonology, working memory, and foreign-language learning. *Q J Exp Psychol A*, 45(1) :21–50, Jul 1992.

Session XX

Psycholinguistique, Cognition, Acquisition

Vendredi 16 juin 2006 - 10h00 12h00

Les effets de compétition lors de la reconnaissance des mots parlés: quand l'inhibition bottom-up joue un rôle.

Sophie Dufour[1], Ronald Peereman[2]

1 Laboratoire de psycholinguistique expérimentale – Université de Genève – 40 Bd du Pont d'Arve, CH-1205 Genève.
2 Laboratoire d'étude de l'apprentissage et du développement – CNRS/Université de Bourgogne – Esplanade Erasme, 21065 Dijon.

sophie.dufour@pse.unige.ch; Ronald.Peereman@u-bourgogne.fr

ABSTRACT

In two experiments, we examined lexical competition effects using the phonological priming paradigm in a repetition task. Experiment 1 showed that inhibitory priming effect occurs when the primes mismatched the targets on the last phoneme (/bagaR/–/bagaʒ/). In contrast, a facilitatory priming effect was observed when the primes mismatched the targets on the medial phoneme (/viRaʒ/-/vilaʒ/). Experiment 2 replicated theses findings with primes presented visually rather than auditorily. The data thus indicate that the position of the mismatching phoneme is a critical factor in determining the competition effect in phonological priming. Such an observation suggests that both bottom-up inhibition and lexical competition are involved in the word recognition process.

1. INTRODUCTION

L'identification des mots présents dans le signal de parole est une étape fondamentale du processus global de compréhension du langage parlé. Il est désormais admis qu'à l'écoute d'un mot, des mots phonologiquement proches du mot entendu sont activés et qu'ils affectent son temps de reconnaissance. Cette influence des compétiteurs est exprimée dans la plupart des modèles actuels de la reconnaissance des mots parlés.

Une façon d'étudier le processus de compétition consiste à mesurer les performances sur un mot cible après la présentation d'un autre mot qui lui est phonologiquement similaire. Le paradigme d'amorçage phonologique s'avère ainsi être un outil intéressant pour étudier la compétition lexicale puisqu'un compétiteur est explicitement présenté et son effet sur le traitement subséquent du mot cible peut être mesuré. Bien que les premières études en amorçage phonologique aient rapporté des résultats contradictoires (cf. [1], pour une revue), des études plus récentes contrôlant le degré de recouvrement entre les amorces et les cibles ainsi que la contribution de facteurs stratégiques ont montré que le temps de reconnaissance d'un mot est ralenti par la présentation préalable d'une amorce qui débute par les mêmes phonèmes (*bouche* – *BOULE*) [1, 2, 3]. Cette observation est compatible avec la plupart des modèles actuels de la reconnaissance des mots parlés qui prédisent que la présentation préalable d'un compétiteur devrait de part l'activation résiduelle qui lui est associée, augmenter son pouvoir de compétition durant le traitement du mot cible. Notons que l'amorce exerce une influence inhibitrice que lorsqu'elle partage ses premiers phonèmes avec le mot cible.

Dans une étude récente, nous avons montré [3] que le degré de désappariement entre les amorces et les cibles est un facteur déterminant pour l'observation d'un effet de compétition en situation d'amorçage phonologique. Plus particulièrement, une inhibition de traitement a été observée lorsque les amorces et les cibles divergent sur le dernier phonème (*bagarre-BAGAGE*) mais pas lorsqu'elles divergent sur les deux derniers phonèmes (*baguette-BAGAGE*). Cette observation est compatible avec le modèle Trace [4] qui postule que le degré de compétition entre deux mots est fonction de leur degré de recouvrement.

Les données que nous venons de présenter indiquent que des inhibitions de traitement sont susceptibles d'être observées à la condition que les amorces et les cibles divergent sur un seul phonème. La présente recherche se situe directement dans le prolongement de notre première étude [3] et examine dans deux expériences si la position du phonème divergent a un impact sur l'amplitude des effets d'amorçage inhibiteurs. La position du phonème divergent peut en effet être déterminante si l'on considère le modèle Shortlist [5] qui intègre des inhibitions bottom-up, en plus d'un mécanisme de compétition entre les mots. Dans le cas d'une divergence précoce entre les amorces et les cibles (*virage–VILLAGE*), l'amorce *virage* devrait décroître très rapidement en activation via une inhibition bottom-up et ceci dès le traitement du phonème /l/ de la cible *VILLAGE*. L'amorce étant rapidement inhibée, elle ne devrait pas agir comme un fort compétiteur du mot cible. Au contraire, dans le cas d'une divergence tardive entre les amorces et les cibles (*bagarre-BAGAGE*), l'amorce *bagarre* devrait continuer de croître en activation jusqu'à ce que le dernier phonème de la cible *BAGAGE* soit traité. Dans une telle condition, l'amorce étant plus fortement réactivée, elle devrait agir comme un plus fort compétiteur du mot cible.

L'effet engendré par la présentation préalable d'une amorce divergeant du mot cible sur le phonème médian (*virage–VILLAGE*) a été comparé à celui obtenu dans le cas d'une divergence sur le dernier phonème (*bagarre-BAGAGE*). Afin de minimiser l'influence de facteurs stratégiques pouvant neutraliser les effets d'amorçage

inhibiteurs [2], les listes expérimentales incluaient une faible proportion d'items reliés (25%). Pour éviter que l'activation résiduelle associée aux amorces se dissipe avant la présentation des mots cibles, un court ISI (50 ms) entre les amorces et les cibles a été utilisé. Certaines études ayant rapporté des effets inhibiteurs de plus grande amplitude lorsque les amorces sont de plus basse fréquence que les mots cibles [1], le mot le moins fréquent des paires amorces et cibles a systématiquement été utilisé en amorce. Dans l'Expérience 1, les amorces et les cibles étaient présentées auditivement. Dans l'Expérience 2, les amorces étaient présentées visuellement et les cibles auditivement. Une tâche de répétition de mots a été utilisée dans chaque expérience.

2. EXPERIENCE 1

2.1. Méthode

2.1.1. Participants

40 sujets de l'Université de Bourgogne ont participé à l'expérience en échange d'un crédit de cours. Tous étaient de langue maternelle française et n'ont rapporté aucun trouble de l'audition ou de la parole.

2.1.2. Matériel

Deux groupes de 28 mots bisyllabiques ont été sélectionnés à partir de Brulex [6]. Dans le premier groupe, les amorces reliées divergeaient des mots cibles sur le dernier phonème (*bagarre-BAGAGE*). Dans le second groupe, les amorces reliées divergeaient des mots cibles sur le phonème médian (*virage–VILLAGE*). Pour chacun des mots, une amorce contrôle n'ayant aucune relation avec le mot cible a été sélectionnée. Les caractéristiques des amorces et des cibles sont présentées dans le Tableau 1. Afin que chaque mot cible soit précédé des deux types d'amorces (reliée, contrôle) et qu'un même sujet ne voie pas deux fois le même mot cible, deux listes expérimentales ont été créées. Chaque liste incluait les 56 mots cibles. La moitié d'entre eux étaient précédés d'une amorce reliée et l'autre moitié d'une amorce contrôle. Les listes ont été contrebalancées de sorte à ce que chaque mot cible soit précédé des deux types d'amorces. Pour atteindre une proportion de paires amorces et cibles reliées de 25%, 56 essais de remplissage sans aucune relation entre les amorces et les cibles ont été rajoutés dans chaque liste.

2.1.3. Procédure

Les stimuli ont été enregistrés par une locutrice de langue maternelle française à l'aide d'un DAT et ont été digitalisés à un taux d'échantillonnage de 44 kHz avec une résolution de 16 bits. Les participants ont été testés individuellement dans une chambre insonorisée. La présentation des items était contrôlée par un Macintosh et les temps de réaction (TRs) ont été récoltés à l'aide d'une clé vocale connectée à l'ordinateur. Les amorces et les cibles étaient présentées dans des écouteurs à un niveau sonore confortable. Un intervalle de 50 ms (ISI) séparait la fin de présentation de l'amorce et le début de présentation du mot cible. Il était demandé aux participants de répéter le plus rapidement et le plus précisément possible le mot cible. La réponse du sujet et le début de présentation de l'amorce de l'essai suivant étaient séparés par un délai de deux secondes. Les latences de répétition étaient mesurées à partir du début de présentation du mot cible jusqu'à la réponse du sujet. Les participants ont été testés sur une seule liste expérimentale et ont commencé l'expérience avec un bloc de 16 essais d'entraînement.

Table 1 : Caractéristiques des amorces et des cibles.

		Amorce	
Divergence Finale	Cible	Reliée	Contrôle
Fréquence[1]	3.47	2.28	2.46
Nb Phon[2]	5	5	5
PU[3]	5.54	5.50	5.50
Durée[4]	618	633	593
Divergence Médiane	Cible	Reliée	Contrôle
Fréquence	3.45	2.35	2.61
Nb Phon	5	5	5
PU	5.61	5.54	5.54
Durée	623	625	612

Notes: [1] en logarithme ; [2] Nombre de Phonèmes ; [3] Point d'Unicité Phonologique ; [4] en millisecondes

2.2. Résultats et Discussion

Pour chaque sujet, les temps de réaction plus grands que 1200 ms et ceux plus grands que 2,5 écart-types au-dessus et en-dessous de leurs temps moyens dans chaque condition ont été exclus des analyses (2.05%). Les temps de réaction obtenus dans chaque condition sont présentés dans le Tableau 2. Les erreurs ayant été peu nombreuses (moins de 1%), les analyses ont été effectuées seulement sur les temps de réaction. Des analyses de variance (ANOVAs) par sujets (F_1) et par items (F_2) ont été conduites avec le type d'amorces (reliée, contrôle) et la position du phonème divergent (médiane, finale) comme variables.

L'effet de la position était significatif seulement par sujets [$F_1(1,39) = 6.87$, $p < .05$; $F_2(1,54) = 0.59$, $p > .20$]. L'effet du type d'amorces était significatif par sujets ($F_1(1,39) = 4.94$, $p < .05$) et approchait la significativité par items ($F_2(1,54) = 3.75$, $p = .06$). L'interaction entre la position et le type d'amorces était significative [$F_1(1,39) = 43.24$, $p < .001$; $F_2(1,54) = 29.97$, $p < .001$].

Des comparaisons planifiées ont été conduites afin de tester l'effet d'amorçage à l'intérieur de chaque position. Un effet d'amorçage inhibiteur a été observé dans le cas d'une divergence finale entre les amorces et les cibles. Les temps de réponse sur les mots cibles étaient en moyenne plus lents de 29 ms lorsqu'ils étaient précédés d'une amorce reliée que lorsqu'ils étaient précédés d'une amorce contrôle [$F_1(1,39) = 45.28$, $p < .001$; $F_2(1,54) = 27.46$, $p < .001$]. Au contraire, un effet d'amorçage facilitateur a été observé dans le cas d'une divergence médiane entre les amorces et les cibles. Les temps de réponse sur les mots cibles étaient en moyenne plus rapides de 17 ms lorsqu'ils

étaient précédés d'une amorce reliée que lorsqu'ils étaient précédés d'une amorce contrôle [$F_1(1,39) = 13.30$, $p < .001$; $F_2(1,54) = 6.26$, $p < .05$].

Table 2 : Temps de réaction moyens (en millisecondes) obtenus dans l' Expérience 1 en fonction de la position du phonème divergent et du type d'amorces (les écart-types sont donnés entre parenthèses)

	Amorce	
Position	Contrôle	Reliée
Finale	812 (102)	841 (96)
Médiane	825 (96)	808 (103)

Comme dans l'étude de Dufour et Peereman [3], un effet de compétition a été observé dans le cas d'une divergence finale entre les amorces et les cibles (*bagarre-BAGAGE*). Par contre, aucun effet de compétition n'a été rapporté lorsque les amorces et les cibles divergent sur le phonème médian (*virage–VILLAGE*). Au contraire, la présentation préalable d'une amorce divergeant du mot cible sur le phonème médian facilite le traitement subséquent du mot cible. Les amorces et les cibles dans le cas d'une divergence médiane se recouvrant sur les phonèmes finaux, l'effet facilitateur pourrait simplement résulter d'un partage de la rime (*virage–VILLAGE*). En effet, plusieurs études montrent un effet facilitateur lorsque les amorces et les cibles diffèrent par les phonèmes initiaux mais partagent les phonèmes finaux (cf. [7] pour une revue). Cet effet disparaissant dans une situation d'inter-modalité, celui-ci a été attribué à des processus qui opèreraient avant l'accès lexical et reflèterait plus particulièrement l'activation de codes pré-lexicaux. Dans le cas où les amorces et les cibles partagent les phonèmes finaux, les mêmes unités en l'occurrence la rime, qui bénéficient d'une activation résiduelle vont être réutilisées ce qui a pour conséquence de faciliter le traitement pré-lexical de la cible. Une possibilité alors est que l'effet inhibiteur dans le cas d'une divergence médiane ait été masqué par une facilitation de nature pré-lexicale.

3. EXPERIENCE 2

Les effets facilitateurs lors d'un recouvrement final disparaissant dans une situation d'inter-modalité, nous avons testé à nouveau l'effet engendré par une divergence sur le phonème médian mais avec une présentation visuelle plutôt qu'auditive des amorces. Une telle manipulation permet en effet, au vu des résultats obtenus dans la littérature, de neutraliser l'effet facilitateur lié au partage d'unités pré-lexicales.

3.1. Méthode

3.1.1. Participants

38 sujets de l'Université de Bourgogne ont participé à l'expérience. Ils ont été recrutés selon les mêmes critères que dans l'Expérience 1.

3.1.2. Matériel

Le matériel était le même que celui utilisé dans l'Expérience 1.

3.1.3. Procédure

La procédure était la même que celle de l'Expérience 1 sauf que les amorces étaient présentées visuellement pendant 350 ms et étaient précédées d'un point de fixation apparaissant au centre de l'écran pendant 500 ms. 50 ms après la fin de présentation de l'amorce, la cible auditive était présentée, conduisant ainsi à un SOA de 400 ms.

3.2. Résultats et Discussion

Les temps de réaction ont été analysés selon les mêmes critères que dans l'Expérience 1. Adoptant ce critère, 0.89 % des données ont été exclues. Les temps de réaction moyens obtenus dans chaque condition sont présentés dans le Tableau 3. Les erreurs ayant été peu nombreuses (moins de 1%), les analyses ont été effectuées seulement sur les temps de réaction. Des ANOVAs ont été conduites avec la position du phonème divergent et le type d'amorces comme facteurs.

Table 3 : Temps de réaction moyens (en millisecondes) obtenus dans l' Expérience 2 en fonction de la position du phonème divergent et du type d'amorces (les écart-types sont donnés entre parenthèses)

	Amorce	
Position	Contrôle	Reliée
Finale	866 (106)	886 (97)
Médiane	859 (94)	849 (95)

L'effet de la position était significatif par sujets ($F_1(1,37) = 36.18$, $p < .001$) mais pas par items ($F_2(1,54) = 2.39$, $p = .13$). L'effet du type d'amorces n'était pas significatif [$F_1(1,37) = 1.77$, $p = .19$; $F_2(1,54) = 1.14$, $p > .20$]. L'interaction entre la position et le type d'amorces était significative [$F_1(1,37) = 10.33$, $p < .01$; $F_2(1,54) = 18.46$, $p < .001$].

Des comparaisons planifiées ont été conduites afin de tester l'effet d'amorçage à l'intérieur de chaque position. Un effet d'amorçage inhibiteur a été observé dans le cas d'une divergence finale entre les amorces et les cibles. Les temps de réponse sur les mots cibles étaient en moyenne plus lents de 20 ms lorsqu'ils étaient précédés d'une amorce reliée que lorsqu'ils étaient précédés d'une amorce contrôle [$F_1(1,37) = 8.18$, $p < .01$; $F_2(1,54) = 14.38$, $p < .001$]. A nouveau, un effet d'amorçage facilitateur a été

observé dans le cas d'une divergence médiane entre les amorces et les cibles. Les temps de réponse sur les mots cibles étaient en moyenne plus rapides de 10 ms lorsqu'ils étaient précédés d'une amorce reliée que lorsqu'ils étaient précédés d'une amorce contrôle [$F_1(1,37) = 4.30$, $p < .05$; $F_2(1,54) = 5.21$, $p < .05$).

En résumé, l'Expérience 2 réplique parfaitement les résultats obtenus dans l'Expérience 1 et indique qu'un effet de compétition n'est observé que dans le cas d'une divergence finale entre les amorces et les cibles.

4. DISCUSSION

Cette étude a été conduite dans le but d'examiner si la position du phonème divergent entre des amorces et des cibles est un facteur déterminant dans l'obtention d'un effet de compétition en amorçage phonologique. Dans l'Expérience 1 comme dans l'Expérience 2, une inhibition de traitement n'a été rapportée que dans le cas d'une divergence finale entre les amorces et les cibles (*bagarre* - *BAGAGE*). Aucune inhibition n'a été mise en évidence lors d'une divergence médiane (*virage– VILLAGE*).

L'observation d'un effet inhibiteur est compatible avec le modèle Trace [4] qui intègre un mécanisme d'inhibition entre les mots. Or dans Trace, les mots sont activés en fonction de leur degré d'appariement avec le signal de parole. Il en résulte qu'un même degré d'évidences perceptives entre le signal de parole et un mot devrait donner lieu à un même degré d'activation. La similarité entre les amorces et les cibles étant de quatre phonèmes sur cinq dans les deux conditions d'amorçage utilisées, des effets inhibiteurs semblent prédits quelle que soit la position du phonème divergent. Toutefois, Trace intégrant la temporalité du signal de parole, un désappariement médian est nécessairement plus pénalisant qu'un désappariement final sur le niveau d'activation. L'amorce étant moins réactivée lors d'une divergence médiane, son influence inhibitrice devrait néanmoins être moindre. Des simulations sont nécessaires pour confirmer cette prédiction.

Le fait qu'aucune inhibition de traitement n'ait été observée lors d'une divergence médiane suggère que les mots sont rapidemment désactivés lorsqu'ils deviennent incompatibles avec l'information reçue. Cette observation est compatible avec le modèle Shortlist [5] qui intègre de l'inhibition bottom-up, en plus d' un mécanisme de compétition entre les mots. Dans le cas d'une divergence médiane, l'amorce *virage* décroit rapidement en activation via une inhibition bottum-up et ceci dès le traitement du phonème /l/ de la cible *VILLAGE*. L'amorce étant rapidement inhibée, elle n'agit pas comme un fort compétiteur du mot cible. Par contre dans le cas d'une divergence finale, l'amorce *bagarre* continue de croître en activation jusqu'à ce que le dernier phonème du mot cible *BAGAGE* soit traité. Celle-ci étant plus fortement réactivée, elle peut agir comme un fort compétiteur du mot cible en ralentissant son temps de reconnaissance.

L'ensemble de nos résultats indique qu'une amorce est susceptible d'agir comme un fort compétiteur à la condition qu'elle diverge du mot cible sur le phonème final. Au contraire, dans le cas d'une divergence sur le phonème médian, un effet facilitateur de l'amorçage a été observé. Cet effet s'est avéré être répliqué dans une situation d'inter-modalité avec des amorces présentées visuellement. Les sujets ayant été conscients des amorces, celles-ci ont pu être codées phonologiquement engendrant ainsi l'effet facilitateur lié au partage de la rime. Cette observation entre en conflit avec les études montrant que l'effet facilitateur lors d'un partage de la rime disparaît dans une situation d'inter-modalité [7]. Toutefois, dans ces études les amorces visuelles étant généralement plus longues de l'ordre de 600ms, les activations pourraient s'être dissipées en raison d'un plus grand délai (SOA) entre les débuts de présentation de l'amorce et de la cible. Les amorces et les cibles étant fortement similaires sur le plan orthographique, une autre possibilité est que l'effet facilitateur résulte de la pré-activation du mot cible durant le traitement des amorces. Le mot cible ayant été préactivé, son traitement ultérieur est facilité. Davantage d'études sont nécessaires pour déterminer l'origine de la facilitation dans le cas d'une divergence médiane entre les amorces et les cibles.

BIBLIOGRAPHIE

[1] M. Radeau, J. Morais and J. Segui. Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance*, 21 :1297-1311, 1995.

[2] M. B. Hamburger and L. M. Slowiaczek. Phonological priming reflects lexical competition. *Psychonomic Bulletin and Review*, 3 :520-525, 1996.

[3] S. Dufour and R. Peereman. Lexical Competition in phonological priming: Assessing the role of phonological match and mismatch lengths between primes and targets. *Memory and Cognition*, 31 :1271-1283, 2003.

[4] J.L. McClelland and J.L. Elman. The TRACE model of speech perception, *Cognitive Psychology*, 18 :1-86, 1986.

[5] D. Norris. SHORTLIST: a connectionist model of continuous speech recognition, *Cognition*, 52 :189-234, 1994.

[6] A. Content, P. Mousty and M. Radeau. Brulex : une base de données lexicales informatisée pour le français écrit et parlé, *L'Année Psychologique*, 90 : 551-566, 1990.

[7] N. Dumay, A. Benraïss, B. Barriol, C. Colin, M. Radeau and M. Besson. Behavioral and electrophysiological study of phonological priming between bisyllabic spoken words. *Journal of Cognitive Neuroscience*, 13 :121-143, 2001.

L'émergence du contrôle segmental au stade du babillage : Une étude acoustique

Mélanie Canault[1, 2], Pascal Perrier[2] & Rudolph Sock[1]
[1]Institut de Phonétique de Strasbourg (IPS)
Équipe d'Accueil 1339 - Linguistique, Langues et Parole (LILPA) - Composante Parole et Cognition
[2]Institut de la Communication Parlée - UMR CNRS 5009 - UMR CNR 5009 INPG & Université Stendhal - Grenoble

ABSTRACT

The aim of this work is to look for evidence that a segmental control of speech production could emerge during babbling from mandible rhythm dominance. Our assumption is that such evidence could be found in temporal modulations of mandibular cycle phases. Acoustic analyses of two subjects between 10 and 15 months of age have revealed that at 10 months, the temporal patterns of speech productions are variable, before becoming more stable and similar to adult temporal patterns at 15 months. These findings are interpreted as consequences of the emergence of a speech specific segmental control guided by the imitation of adult productions.

1 INTRODUCTION

Notre recherche s'inspire de la théorie *Frame then Content* de MacNeilage [6], qui défend l'unité articulatoire et rythmique de la structure syllabique au stade du babillage, et l'associe au cycle de l'oscillation mandibulaire. Nous nous appuyons aussi sur l'hypothèse d'une représentation segmentale phonémique du contrôle de la parole chez l'adulte [9]. L'objectif de ce travail est donc d'étudier de quelle manière et à quel moment la structure articulatoire et rythmique unique présente au stade du babillage va se dissocier en ses composantes segmentales vocaliques et consonantiques. Notre hypothèse envisage la variation temporelle du cadre syllabique, que nous supposerons initialement régulier ([5], [1]), comme indice de l'émergence d'un contrôle indépendant du segment. Nous avons donc mesuré les patrons temporels de données *acoustiques* interprétables en termes articulatoires. Plus spécifiquement, nous avons vérifié si le contrôle supposé du segment pouvait émerger graduellement du contrôle temporel indépendant des phases d'ouverture et de fermeture du conduit vocal.

2 LE BABILLAGE, LA SYLLABE ET LE ROLE DETERMINANT DE LA MANDIBULE

2.1 L'oscillation mandibulaire comme cadre de l'émergence de la parole

Le babillage est considéré comme le stade du développement langagier au cours duquel les premières syllabes émergent. Selon l'hypothèse de MacNeilage [6], les éléments constitutifs de la syllabe forment une unité articulatoire, déterminée par des contraintes physiologiques. Les articulations les plus précoces seraient ainsi ordonnées par le mouvement mandibulaire dont la cyclicité, due à l'alternance des phases d'ouverture et de fermeture, suffirait à l'émergence du cadre syllabique. La dominance du cycle mandibulaire imposerait alors l'absence d'un contrôle indépendant des autres articulateurs. Munhall et Jones [8] ont notamment confirmé cette hypothèse en montrant, à travers une étude cinématique réalisée chez un sujet de 8 mois, la non implication de la lèvre supérieure dans le mouvement de fermeture de la cavité buccale.

L'oscillation mandibulaire jouerait donc un rôle majeur dans l'organisation articulatoire du babillage et pourrait être considérée comme le support des aménagements articulatoires à venir. En effet, les stratégies de contrôle de la mandibule atteindraient leur maturité les premières. Ainsi, dans une étude cinématique comparative des déplacements verticaux de la mandibule et des lèvres, menée chez des adultes et des enfants (1, 2, 6 ans), Green *et al* [4] ont montré que chaque articulateur avait un processus développemental unique, et que la mandibule accédait à un patron de mouvement proche de celui de l'adulte plus tôt que les lèvres. Pour ces auteurs, les structures articulatoires pourvues d'un degré de liberté plus grand nécessiteraient un processus d'apprentissage du contrôle plus long. Ainsi la langue et les lèvres, étant déformables, représenteraient des systèmes plus complexes à contrôler. Green *et al* [3] ont également étudié la coordination mandibule/lèvre chez des sujets du même âge, à travers l'observation du couplage temporel et spatial de leur déplacement au cours de séquences Consonne bilabiale-Voyelle. Les résultats montrent que la contribution de la mandibule à la fermeture orale est très forte à 1 an et diminue à 2 ans. En revanche, celle des lèvres augmente entre 2 et 6 ans. Le fort couplage spatial et temporel des lèvres avec la mandibule reflète l'absence d'un contrôle indépendant à 1 an.

Selon Davis et MacNeilage [1], le cadre syllabique constituerait la structure temporelle de base au sein de laquelle les éléments du contenu vont se développer grâce à l'acquisition du contrôle indépendant des articulateurs. L'émergence du contrôle autonome des articulateurs se traduira par l'organisation coordonnée, spatialement et temporellement, des gestes articulatoires. De cette manière, consonne (C) et voyelle (V) émergeront graduellement comme des entités indépendamment contrôlables au sein de la syllabe.

2.2 Perturbation du cycle oscillatoire et émergence du segment

Le rôle de la mandibule pèse sur l'organisation structurelle, mais aussi temporelle des premiers énoncés. Ces derniers sont souvent décrits comme des énoncés redupliqués du type /bababa…/, dont les séquences syllabiques sont perceptuellement isochrones. La fermeture consonantique, associée à un signal acoustique de faible énergie et de courte durée, et l'ouverture vocalique, associée à un signal acoustique de forte énergie et d'une durée plus longue, produisent l'effet d'une régularité temporelle des séquences CV [1]. Ce cycle oscillatoire constitue le cadre temporel s de la structure syllabique [2], qui serait alors dépourvue de toute

organisation segmentale. Dans ce contexte, il est légitime de faire l'hypothèse que la perturbation du cadre syllabique pourrait constituer les premiers indices de l'émergence d'un contrôle segmental. Le premier pas vers ce contrôle passerait par la maîtrise d'événements temporels nouveaux, dont la manifestation serait la variation de la durée des syllabes successives. Dans ce contexte, nous prédisons que, au cours de l'émergence du contrôle segmental, après les énoncés canoniques du type /bababa…/, le bébé passera à des énoncés du type /babaababaaaba…/ ayant perdu leur régularité temporelle. Plus encore, la variation temporelle des phases constitutives du cycle mandibulaire engendrera la désolidarisation des éléments du cadre et par conséquent leur indépendance ; d'où l'émergence du contrôle segmental.

3 METHODE

3.1 Les données

Sujets

Deux enfants de sexe masculin, âgés de 9 mois et de 8 mois, ont été recrutés, sur autorisation parentale, dans une crèche strasbourgeoise. Aucun d'entre eux ne présentait de troubles moteurs, auditifs ou psychologiques.

Collecte de données

Des enregistrements audio individuels, d'une durée de 20-25 minutes, ont été réalisés chaque semaine pendant 6 mois. Le sujet était alors installé dans un parc dans le but de réduire son espace de déplacement. Ses jouets favoris étaient mis à sa disposition afin d'initier une phase ludique d'interaction avec l'investigateur. Même si les premières séances furent moins fructueuses, la phase d'habituation ne fut que de courte durée. Le système d'acquisition était constitué d'un DAT (Sony TCD-D3) et d'un micro directif (BST à condensateur UM-3). Il était placé à proximité du bébé, mais hors de sa portée afin de ne pas trop attirer son attention.

3.2 Etiquetage du signal acoustique

Environ 1 à 2 minutes de productions enfantines pouvaient, en moyenne, être extraites d'une séance d'enregistrement. Sur ces quelques minutes nous avons sélectionné les séquences exploitables en fonction de leur type ; nous avons retenu les types Cocclusive-V ainsi que V-Cocclusive-V, et éliminé les séquences trop bruitées. Par la suite, chacune des séquences babillées fut transcrite, puis étiquetée à l'aide du logiciel Praat®. Nous avons porté notre intérêt sur les événements cycliques au sein des énoncés polysyllabiques afin de rendre compte de l'existence et de l'évolution de la régularité temporelle syllabique défendue par MacNeilage. Notons que les énoncés bisyllabiques se sont révélés majoritaires. Un cycle peut être défini comme l'intervalle temporel existant entre la réitération d'un même événement acoustique. Ainsi, avons-nous désigné dans un premier temps comme cycle, la distance temporelle existant entre deux relâchements (cycle 1 ou le cycle des relâchements) lorsque la séquence est constituée de 2 syllabes CV au minimum.

Malheureusement, la fiabilité de cette segmentation au sein d'un signal bruité est restreinte à cause d'une explosion-friction difficilement repérable. C'est pour cette raison que nous avons préféré exclure ce cycle et en considérer un autre, le cycle 2 ou le cycle vocalique. Celui-ci s'étend du début d'une structure formantique vocalique stable au début de la structure formantique vocalique stable subséquente (figure 1). En termes articulatoires, ce cycle s'initie par un état suffisamment ouvert du conduit vocal permettant l'apparition des résonances vocaliques adéquates ; il encadre la phase consonantique obstruente.

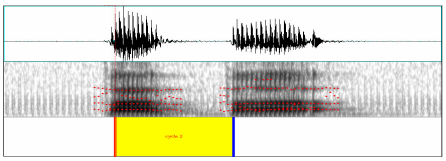


Figure 1. Détermination des bornes du cycle 2

Cependant, nous nous sommes rendus à l'évidence qu'il était impossible, en utilisant ce procédé, de prendre en compte la durée de la voyelle finale dans le cadre de ce cycle. Par conséquent, nous avons choisi d'intégrer un troisième marquage cyclique, que nous appellerons cycle 3 ou le cycle consonantique, qui va de la fin d'une structure formantique stable à la fin de la structure formantique stable suivante (figure 2). C'est un cycle de closion articulatoire, englobant la phase vocalique V2.

Notre analyse portera donc sur une étude systématique des cycles 2 et 3, même si nous ne livrerons ici que des résultats pour le cycle 3. Les tendances observées pour ces deux cycles sont en effet similaires, mais elles sont plus nettes pour le cycle 3.

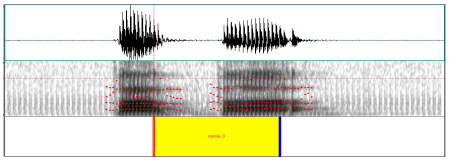


Figure 2. Détermination des bornes du cycle 3

3.3 Analyse

Groupe d'âge

Nous avons établi des classes d'âge bien distinctes en centrant notre intérêt sur trois étapes du développement : 9-10 mois; 12 mois et 14-15 mois. Ces étapes correspondent respectivement : aux stades du babillage, à la fin du babillage et au début des premiers mots, et enfin à celui des mots.

Projections graphiques

La structure même de nos données expérimentales, caractérisées en particulier par un nombre très hétérogène de mesures et par des variances très différentes selon les catégories d'âge, ne nous autorise pas à pratiquer une analyse statistique classique sous forme de test de Student ou d'analyse de la variance. Aussi, afin de rendre compte de l'émergence de l'indépendance segmentale, avons nous eu recours, au moins dans une première phase de notre analyse, à l'observation graphique de l'évolution des différentes caractéristiques temporelles au sein des cycles. En premier lieu, nous avons observé l'évolution de la proportion vocalique au sein du cycle 3. Cette phase d'ouverture pourrait en effet porter les premières traces de variation. Il semble, en effet, plus évident de contrôler la configuration buccale ouverte que la configuration fermée. Contrairement à la phase de fermeture, l'ouverture n'est pas contrainte par les mêmes pressions aériennes supra glottiques. Elle pourrait, par conséquent, subir de plus grands changements temporels.

Puis, nous nous sommes orientés vers l'étude du rapport existant entre les valeurs des durées consonantiques et des durées vocaliques dans le cycle. D'une part, pour examiner l'existence éventuelle d'une variation simultanée des deux phases du cycle et d'autre part, pour savoir si une relation entre l'évolution de leur variation pourrait être mise au jour.

4 RESULTATS

4.1 Evolution du cycle et de la proportion vocalique

La proportion vocalique est déterminée par la valeur que la structure formantique stable prend au sein d'un cycle donné.

Sujet 1

Les ellipses de dispersion (à 2σ) des durées vocaliques et cycliques (figure 3) indiquent une certaine variabilité à 10 mois. La durée du cycle 3 s'étend de 194.3 ms à 989 ms (valeur moyenne # 500 ms) et le pourcentage de la phase vocalique varie entre 44.7% et 84, 69%. Puis, à 12 mois, la durée du cycle tend à diminuer (valeur moyenne # 350 ms) et à se stabiliser, sa variabilité décroissant sensiblement, tandis qu'à l'intérieur du cycle, la variabilité de la proportion vocalique se maintient. Enfin, à 15 mois, la durée cyclique se centralise autour des 300 ms s'approchant ainsi de la durée moyenne d'une syllabe adulte et de sa période d'oscillation préférentielle qui se situe autour de 3Hz [7], tandis que sa variabilité se réduit encore de manière sensible. Les proportions vocaliques se concentrent aux environs de 40% du cycle avec une variabilité modérée par rapport aux deux tranches d'âge précédentes.

Sujet 2

Contrairement au sujet 1, le sujet 2 présente un rythme mandibulaire moyen avoisinant les 3Hz et de variabilité réduite dès les premiers enregistrements, c'est-à-dire à un stade très précoce (figure 4). Dès 10 mois, la variation cyclique se concentre entre 200 ms et 400 ms (soit 2-4 Hz). La proportion vocalique moyenne au sein du cycle est stable dans la période d'âge analysée : elle se situe, comme pour le sujet précédant, autour de 40%. En comparaison avec le sujet 1, on observe que la variabilité de la proportion vocalique à 10 et 12 mois est visiblement restreinte, mais qu'à 14 mois elle est similaire.

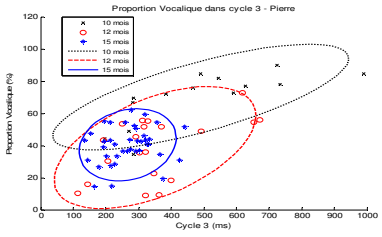


Figure 3. Durée du cycle 3 (ms) et V (%) - sujet 1

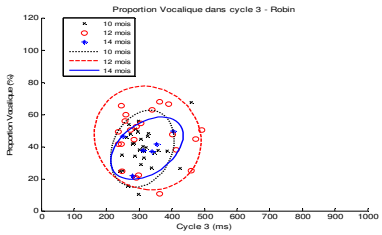


Figure 4. Durée du cycle 3 (ms) et V (%) - sujet 2

4.2 Relation entre durée consonantique et durée vocalique

La durée consonantique est déduite à partir de la durée de V au sein du cycle.

Sujet 1 (Figure 5)

A dix mois, c'est la phase vocalique qui est la plus variable. Dans la tranche d'âge suivante, la variabilité vocalique diminue tandis qu'elle augmente pour la consonne. La valeur moyenne de la durée vocalique décroît sensiblement, la durée consonantique moyenne restant sensiblement constante. Enfin, à 15 mois on observe non seulement une réduction notoire de la variabilité de la durée vocalique au sein du cycle, mais aussi une tendance à l'équilibre des proportions des phases vocalique et consonantique. En effet, le centre de l'ellipse de dispersion se rapproche du point où les deux durées sont égales (entre 100-200 ms : figure 5).

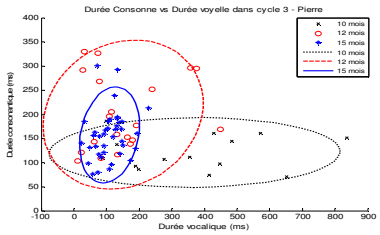


Figure 5. Durées consonantique et vocalique - sujet 1

Sujet 2 (Figure 6)

La variabilité est dès lors présente à 10 mois, mais elle touche essentiellement à la consonne, à l'inverse du sujet 1. Le sujet 2, favorise donc, au cours de la phase d'exploration temporelle, la modulation de la phase de fermeture du cycle mandibulaire. Puis à 12 mois, la variabilité augmente. Nous avions relevé que le sujet 1, au même stade, réduisait sa variabilité vocalique et amplifiait sa variation consonantique. Ce qui est intéressant c'est le fait que, malgré cette divergence, la consonne occupe un espace de dispersion relativement proche, à 12 mois, chez les deux sujets (100 ms à 300 ms). Enfin, à 14 mois une réduction de la variabilité globale semble s'opérer pour ce sujet. Dans la mesure où le nombre d'occurrences reste trop faible pour interpréter ces résultats comme étant robustes, il convient de considérer ces résultats avec prudence. Notons cependant, que ce phénomène est conforme aux observations faites pour le sujet 1.

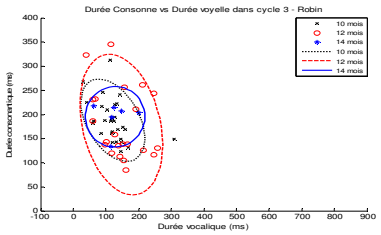


Figure 6. Durées consonantique et vocalique - sujet 2

5 DISCUSSION ET INTERPRÉTATION

A première vue, la conclusion qui s'imposerait à la lecture de nos résultats serait que nos mesures ne vont pas totalement dans le sens de nos hypothèses. Nous avions prédit une progression des perturbations de la régularité temporelle, attestée par MacNeilage au stade précoce du babillage, qui traduirait l'émergence d'un contrôle à l'intérieur du cycle se caractérisant par une variabilité croissante de l'organisation temporelle du cadre et de son contenu. Or, c'est bien le constat inverse que nous faisons tout particulièrement pour le sujet 1. Cependant, une étude plus attentive permet de tempérer cette première analyse et de préciser nos hypothèses en les confirmant. En effet, nous relevons que, même si leurs productions évoluent de manières très différentes entre 10 et 15 mois, les deux sujets étudiés convergent, lors de la dernière période d'âge étudiée, vers des caractéristiques très similaires : un cycle mandibulaire de fréquence moyenne 3Hz, une proportion de la phase vocalique située entre 40 et 45%, et une variabilité temporelle réduite tant pour le cycle que ses composantes. La fréquence 3Hz étant la fréquence préférentielle des oscillations mandibulaires de l'adulte [7], on peut faire l'hypothèse que les productions des deux bébés évoluent vers les patrons temporels qui sont ceux de l'adulte.

D'autre part, le sujet 1 montre à 10 mois une très grande variabilité, d'abord dans la phase vocalique, puis dans la phase consonantique. Ceci peut-être considéré comme cohérent avec nos prédictions si on fait l'hypothèse (justifiée si on se réfère à la littérature du domaine) que ce bébé n'est plus au début de la phase de babillage canonique, traditionnellement située autour de 6 mois. Il aurait donc déjà dépassé la phase initiale du cycle mandibulaire intrinsèquement régulier décrite par MacNeilage. Le stade d'irrégularité et d'apprentissage d'un contrôle spécifique au sein du cycle mandibulaire des phases d'ouverture et de fermeture, que nos hypothèses laissaient prévoir, serait déjà amorcé. Si on accepte cette perspective, qu'il conviendra cependant de vérifier par d'autres données expérimentales à un âge plus précoce, nous pouvons supposer que le sujet 2 avait déjà atteint dès les premiers enregistrements un stade plus avancé que le sujet 1. Il serait déjà dans la deuxième phase de cet apprentissage segmental, celui de l'affinement du contrôle intégrant plus de régularité pour une évolution vers les productions de l'adulte.

Cette deuxième phase constitue la manifestation la plus probante de l'émergence du contrôle dans nos données, car l'évolution du cycle mandibulaire observé chez le sujet 1, passant d'une fréquence de 5 à 6 Hz à la fréquence préférentielle de l'adulte, ne peut pas s'expliquer par des critères purement physiques. La croissance crânienne de l'enfant entre 8 et 15 mois ne saurait en effet justifier une telle variation de la fréquence modale de la mandibule. Il s'agit donc bien d'une variation contrôlée vraisemblablement guidée par le mimétisme à celui de l'adulte [10].

6 CONCLUSIONS

En conclusion, nos résultats peuvent être interprétés dans le cadre de nos hypothèses sur l'émergence d'un contrôle segmental au cours de la période allant du babillage tardif aux premiers mots. Dans une première phase, le bébé se familiariserait avec les possibilités de variation temporelle du cycle mandibulaire et des éléments de son contenu, avant d'affiner, dans une seconde phase, le contrôle du timing de ses mouvements. La première phase, marquée par une large variabilité temporelle (10 mois), serait le reflet de l'émergence d'un contrôle indépendant des éléments articulatoires au sein du cadre syllabique. Puis, viendrait la phase de réduction de la variabilité (14-15 mois), laquelle impliquerait le contrôle plus précis des mouvements et cela par mimétisme avec les productions de l'adulte [10]. Par la collecte de données à un stade plus précoce, nos futurs travaux viseront à consolider ce cadre d'hypothèses.

Remerciements à ACI TTT 2003-2006, MER

7 BIBLIOGRAPHIE

[1] B.L. Davis and P.F. MacNeilage. The articulatory basis of babbling. In *Journal of Speech and Hearing Research,* 38:1199-1211, 1995.

[2] B.L. Davis and P.F. MacNeilage. Organisation of babbling: a case study. In *Language and Speech,* volume 37:341-355, 1994.

[3] J.R. Green, C.A. Moore and K.J. Reilly. The sequential developpment of jaw and lip control for speech. In *Journal of Speech, Language, and Hearing Research,* 45: 66-79, 2002.

[4] J.W. Green, C.A. Moore, M. Higashikawa and R.W. Steeve. The physiologic development of speech motor control: lip and jaw coordination. In *Journal of Speech, Language, and Hearing Research,* 43: 239-255, 2000.

[5] G. Konopczynski. Vers un modèle développemental du rythme français : problèmes d'isochronie reconsidérés à la lumière des données de l'acquisition du langage. In *Bulletin de l'Institut de Phonétique de Grenoble*, volume 15, pages 157-190, 1986.

[6] P.F. MacNeilage. The Frame/Content theory of evolution of speech production. In *Behavioral and Brain Sciences,* 21:499-546, 1998.

[7] T. Morimoto, T. Inoue, T. Nakamura, T. and Y. Kawamura. Frequency dependent modulation of rhythmic human jaw movements. In *Journal of Dental Research,* 68:1310-1314, 1984.

[8] K.G. Munhall and J.A. Jones. Articulatory evidence for syllabic structure. In *Behavioral and Brain Sciences,* 21:524-525, 1998.

[9] P. Perrier, Y. Payan and R. Marret (2004). Modéliser le physique pour comprendre le contrôle : le cas de l'anticipation en production de parole. In *L'anticipation à l'horizon du présent* (R. Sock & B. Vaxelaire, editors). Sprimont, Belgique : Pierre Mardaga, pages 159-177, 2004.

[10] M. Studdert-Kennedy. Imitation and the emergence of segments. In *Phonetica*, 57:275-283, 2000.

L'implication des contraintes motrices dans « l'effet Labial Coronal »

Amélie Rochet-Capellan et Jean-Luc Schwartz

Institut de la Communication Parlée
INPG / Université Stendhal/ CNRS UMR 5009, Grenoble, France
Amelie.rochet-capellan@icp.inpg.fr jean-luc.schwartz@icp.inpg.fr

ABSTRACT

Stability of L_aC_o (Labial-Coronal) and C_oL_a CVCV sequences was compared using the paradigm of reiterant speech with rate increase. The rationale was that speed would lead the articulatory system towards its most stable coordination mode. A first study analyzed the acoustic productions of 28 French speakers. Then, a second study focused on the articulatory coordination for 5 speakers. Results show that the repetition of L_aC_o and C_oL_a disyllables could both evolve towards a L_aC_o (/pata/->/patá->/ptá) or a C_oL_a (/pata/->/páta->/tpá) attractor. Yet, the L_aC_o attractor is largely favored. Moreover, speed drives lips and tongue occlusions close together on a single jaw cycle. This provides new elements to explain the "LC effect" in world languages by motor control constraints.

1. INTRODUCTION

L'effet LC (Labial-Coronal) réfère au fait que les lexiques des langues contiennent environ 2.5 fois plus de CV.CV et CVC de type L_aC_o (/pata/) que C_oL_a (/tapa/) [1][2]. Cette dissymétrie émergerait avec les premiers mots [3][4]. La principale explication de ce phénomène associe l'hypothèse selon laquelle les occlusions labiales (OL_a) seraient plus faciles à produire que les occlusions coronales (OC_o) et la tendance des systèmes moteurs à initialiser les séquences d'actions par l'action la plus simple [3]. Cette proposition repose sur la théorie « Frame then Content » stipulant que seules les oscillations verticales de la mandibule (le « frame ») seraient contrôlées dans le babillage précoce. La maîtrise des articulateurs portés (le « Content ») apparaîtrait ultérieurement [3][5]. Ainsi, les OL_a résulteraient d'un geste de mâchoire (« pure frames ») alors que les OC_o impliqueraient la superposition d'un geste de la langue. Cependant, cette hypothèse « simple first » se base sur des données développementales et n'a jamais été évaluée expérimentalement. D'autre part, les oscillations mandibulaires peuvent induire aussi bien des OC_o que des OL_a selon la morphologie du bébé et/ou la position statique de la langue [6]. De plus, chez l'adulte, les lèvres sont actives pour les OL_a [7]. Ces données remettent en cause la plus grande simplicité des OL_a et supposent des processus de coordination différents pour l'adulte par rapport au bébé.

Ces processus sont investis ici afin d'expliquer l'effet LC en termes de coordination motrice et d'économie d'énergie.

2. COORDINATION ET STABILITE

L'hypothèse « simple first » repose sur l'idée que les langues favorisent les formes faciles à produire et à entendre [8]. Or, indépendamment de tout axe développemental, l'étude des systèmes moteurs montre que certaines coordinations sont plus stables et plus économiques que d'autres. Ainsi, les chevaux adaptent leur allure à la vitesse de façon à réduire leur consommation d'énergie [9]. Des résultats analogues caractérisent la coordination bimanuelle [10]. La vitesse amènerait donc le système vers son état de coordination le plus stable. Ce constat a été utilisé en parole pour étudier la stabilité de certaines formes relativement à d'autres [11]. Ces études ont notamment montré que la répétition d'une syllabe VC évolue vers une forme CV avec l'augmentation du débit. Cette stabilité articulatoire de CV pourrait expliquer sa dominance dans les langues [12][13]. Ce cadre de recherche apporte deux idées intéressantes : (1) La répétition accélérée d'une série de gestes peut réorganiser la coordination entre les articulateurs vers un mode préférentiel. (2) Les différentes synergies réalisables par un groupe d'articulateurs se caractérisent par des relations de phase [11].

3. HYPOTHESE ALTERNATIVE

Le travail rapporté ici suppose une coordination L_aC_o plus stable et donc plus économique que C_oL_a. Ainsi, la répétition accélérée de CVCV L_aC_o et C_oL_a devrait évoluer vers une forme L_aC_o du fait de deux phénomènes. (1) Les oscillations mandibulaires saturent en deçà d'un certain débit [14]. (2) L'anticipation de OC_o dans OL_a serait meilleure que l'inverse [15], favorisant un rapprochement L_aC_o plutôt que C_oL_a. Ainsi, l'accélération devrait induire une progression d'un cycle de mâchoire par syllabe à un cycle par bisyllabe avec OL_a qui précède OC_o. Ce processus est mesurable par deux indicateurs : (1) En acoustique, la voyelle après OL_a devrait se réduire jusqu'à disparaître ; (2) En articulatoire, la durée entre OL_a et OC_o suivante devrait être plus faible qu'entre OC_o et OL_a suivante. Enfin, la coordination constricteurs/mâchoire devrait évoluer pour permettre la production sur un seul cycle de mâchoire.

Deux études utilisant le paradigme de répétition accélérée ont testé ces hypothèses. La première investit la stabilité des formes L_aC_o et C_oL_a sur la base de données acoustiques. La deuxième décrit les processus articulatoires impliqués.

4. ETUDE ACOUSTIQUE

4.1. Procédure

Les 28 participants avaient le français pour langue maternelle et n'étaient pas informés des objectifs de l'étude. La tâche consistait à accélérer puis décélérer des CVCV L_aC_o (/pata/, /pasa/, /fata/) et C_oL_a (/tapa/, /sapa/, /tafa/) au rythme d'un carré clignotant (noir-blanc-noir-…). Un programme informatique contrôlait l'affichage du carré au centre d'un écran et l'enregistrement du son (16 kHz). La durée du carré

débutait à 300 ms et diminuait de manière linéaire pour atteindre 125 ms au bout de 4 s et 50 ms au bout de 8 s. Elle ré-augmentait ensuite symétriquement pour la décélération. Le but n'était pas de contrôler précisément le débit mais d'homogénéiser la progression et d'amener le locuteur à un débit maximal. La consigne insistait sur la nécessité d'aller vite et de poursuivre en dépit de toutes transformations perçues. Trois listes arrangeant les 6 CVCV dans des ordres aléatoires et différents pour chaque locuteur étaient passées successivement après trois items d'entraînement. Les CVCV s'affichaient à l'écran et le locuteur initialisait le flash en appuyant sur la barre d'espace.

4.2. Hypothèses et mesures

La meilleure coordination L_aC_o par rapport à C_oL_a devrait induire une réduction de la voyelle suivant la consonne labiale jusqu'à sa disparition (/pata/→/patá/→/ptá/). Cette progression peut s'étudier en comparant l'intensité des deux voyelles. La figure 1 représente l'étiquetage des courbes d'intensité des données lors d'une progression vers une syllabe CCV. Les consonnes labiales (C_L) et coronale (C_C) sont repérées aux minima d'énergie et les voyelles (après la labiale : V_L et après la coronale : V_C), aux maxima. Chaque CVCV se caractérise par : (1) Sa durée : délai entre sa première consonne et celle du CVCV suivant. (2) La différence d'intensité entre V_C et V_L (ΔI).

ΔI devrait être positive pour les CVCV L_aC_o et C_oL_a.

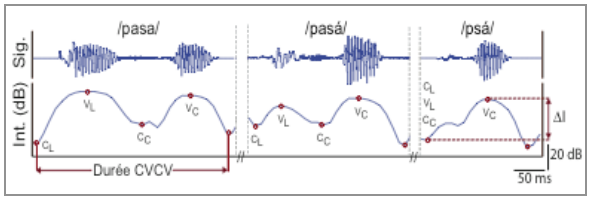


Figure 1 : Etiquetage des courbes d'intensité. Signal acoustique (en haut) et intensité (en bas) en fonction du temps (voir texte pour le détail).

4.3. Résultats

Deux locuteurs ont été exclus de l'analyse pour des problèmes techniques.

Analyse globale

La figure 2 représente ΔI en fonction de la durée des CVCV pour toutes les productions /pasa/-/sapa/ de tous les locuteurs (les résultats étant similaires pour les trois paires L_aC_o / C_oL_a). Les ΔI sont proches de zéro pour les durées supérieures à 250-300 ms. En deçà, l'intervalle de valeurs augmente avec plus de valeurs positives que négatives. Ce pattern indique une progression vers un attracteur L_aC_o (cf. /sápa/, /psá/). Cependant des valeurs négatives s'observent, particulièrement pour les C_oL_a, témoignant d'une progression vers un attracteur C_oL_a (cf. /sapá/, /spá/),

L'analyse a ensuite été restreinte aux productions de durée en deçà de 300 ms.

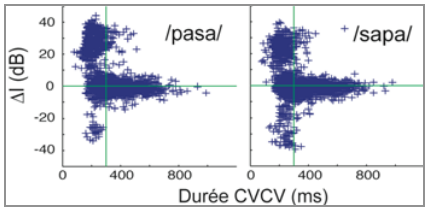


Figure 2 : ΔI en fonction de la durée des CVCV pour toutes les productions /pasa/ et /sapa/ de tous les locuteurs.

Analyse des productions rapides

Afin d'homogénéiser les groupes, l'analyse a été limitée aux sujets ayant au moins 5 productions (CVCV ou CCV) de durée inférieure à 300 ms pour deux séries de chacune des 6 séquences. 21 locuteurs ont été ainsi sélectionnés. Seules les deux séries avec le plus de productions ont été analysées. La moyenne des ΔI des 42 séries est positive pour les 6 séquences et diffère significativement de 0 sauf pour /sapa/ (tableau 1, ligne 1). La répartition globale des moyennes confirme ce résultat (lignes 2 et 4). Enfin, le ratio positives/négatives des moyennes qui diffèrent significativement de 0 va de 2.6 (/tapa/) à 10.5 (/fata/) (lignes 3 et 5) .

Tableau 1 : ΔI pour les productions rapides : moyennes et répartition des moyennes positives et négatives par séquence.

		Pata	Tapa	Pasa	Sapa	Fata	Tafa
Moy.		6,6 **	7,0 **	7,5 **	3,8	8,9**	8,6 **
Moy. pos.	**Tot**	30	30	27	21	37	33
	Sign.	16	13	16	13	21	18
Moy. neg.	**Tot.**	12	12	15	21	5	9
	Sign.	3	5	2	4	2	4

- ** Moyenne significativement différente de 0, $p < 0.01$ / 6, ttest bilatéral
- tot. : toutes les moyennes, sign. : moyennes différant de 0, $p < 0.05$ / (42 * 6), ttest bilatéral

Ainsi, la majorité des répétitions vont vers la forme L_aC_o. La variabilité entre les locuteurs (non détaillée ici) confirme ces tendances avec trois grand profils : (1) évolution systématique vers L_aC_o quelque soit le CVCV; (2) évolution systématique vers C_oL_a et (3) évolution bistable, qui change d'une répétition à une autre ou selon le CVCV. Cependant, le groupe L_aC_o domine largement.

Le but de la deuxième étude était de comprendre les phénomènes articulatoires sous-tendant cette stabilité L_aC_o.

5. ETUDE ARTICULATOIRE

5.1. Procédure

5 locuteurs ont participé à cette étude avec les mêmes critères de sélection que pour l'étude précédente. Le matériel comprenait les 6 CVCV de la première étude et deux items de contrôle : /papa/ et /tata/. La tâche et les consignes de répétition étaient identiques mais sans métronome. Pour chaque essai, l'expérimentateur énonçait le CVCV et le locuteur le répétait en accélérant puis en décélérant. La durée d'enregistrement variait de 10 à 16 s. Trois listes construites selon les modalités de l'expérience 1 étaient passées successivement avec une pause entre deux.

Les trajectoires de la pointe de la langue (PL), de la lèvre inférieure (LI) et de la mâchoire étaient enregistrées avec un

EMA échantillonnant à 500 Hz. Un micro fixé au casque enregistrait le son en parallèle, numérisé ensuite à 20 kHz.

5.2. Hypothèses

A débit lent, chaque constriction devrait se réaliser sur un cycle de mâchoire. Puis, la durée des cycles de mâchoire devrait saturer à une valeur plancher [14]. Associé à l'anticipation et à la meilleure stabilité du geste L_aC_o, ce phénomène devrait induire pour les CVCV L_aC_o et C_oL_a : (1) une bascule sur un seul cycle de mâchoire ; (2) une réorganisation de la coordination entre les constricteurs et la mâchoire et (3) un intervalle de temps plus court de OL_a à OC_o que de OC_o à OL_a. De plus, si la bascule sur un cycle dépend de l'anticipation, elle ne devrait pas s'observer quand le même organe réalise les deux constrictions. Ainsi, les durées des productions devraient être plus longues pour /papa/ et /tata/ que pour les L_aC_o / C_oL_a.

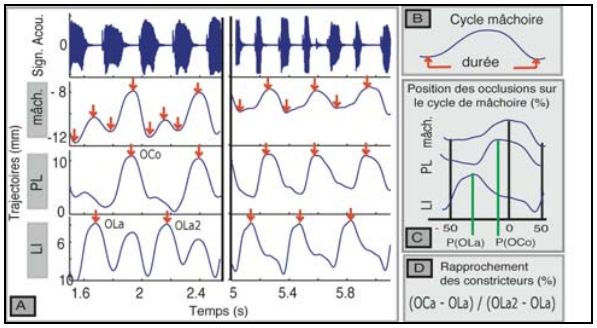


Figure 3 :Etiquetage des trajectoires (A), cycle de la mâchoire (B); mesure du phasage entre OL_a, OC_o et la mâchoire (C) et phasage entre les constricteurs (D).

5.3. Etiquetage et mesures

Les données acoustiques ont été traitées comme dans la première étude. Cet étiquetage a été utilisé pour marquer les trajectoires des constricteurs. Après filtrage (passe bas, Chebychev), les OL_a et les OC_o ont été repérées comme la plus haute position de LI entre V_L et V_C et la plus haute position de PL entre V_C et V_L. Pour les trajectoires de mâchoire, les maxima et les minima ont été repérés automatiquement puis validés manuellement (figure 3, A). Ces étiquetages ont permis de mesurer : (1) la durée des cycles de la mâchoire (figure 3, B); (2) la position de OL_a ($P_{(OLa)}$) et de OC_o ($P_{(OCo)}$) relativement au cycle de la mâchoire (figure 3, C) et (3) la durée de OL_a à OC_o par rapport à celle de OL_a à OLa_2 (figure 3, D).

5.4. Résultats

Bascule sur un seul cycle de mâchoire

Sur l'exemple de la figure 3.A, OL_a et OC_o sont d'abord réalisées sur deux cycles de mâchoire (à gauche) puis sur un seul (à droite). L'étude globale de la durée de cycles de mâchoire selon les durées acoustiques de CVCV généralise la bascule sur un seul cycle. En effet, pour les L_aC_o/C_oL_a (figure 4, en haut à droite), les productions s'agglutinent autour de deux droites : (y = 1/2y) et (y = x). Cette double répartition montre que les CVCV de durées inférieures à environ 400 ms peuvent être réalisés sur deux ou sur un seul cycle de mâchoire. En revanche, les CVCV dupliqués requièrent toujours deux cycles de mâchoire (figure 4, en haut à gauche). D'autre part, les durées de CVCV vont de 200 à 600 ms avec un pic à 250 ms pour les L_aC_o/C_oL_a et à 300 ms pour les dupliqués (figure 4, milieu). Enfin, les durées de cycles de mâchoire vont de 100 à 400 ms avec un pic à 150 ms pour les dupliqués et à 200 ms pour les L_aC_o/C_oL_a (figure 4, en bas). Ainsi, le passage sur un seul cycle permet des débits plus rapides pour des vitesses de mâchoire plus lentes.

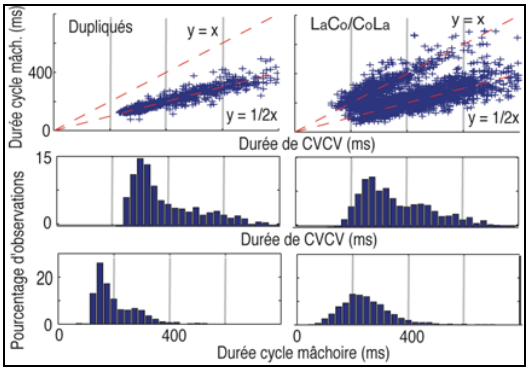


Figure 4 : Pour toutes les productions de tous les locuteurs et pour les CVCV dupliqués (à gauche) et L_aC_o/C_oL_a (à droite) : durée des cycles de mâchoire en fonction de la durée des CVCV (en haut); répartition des durées de CVCV (au milieu) et des durées de cycles de mâchoire (en bas).

Phasage entre la mâchoire et les constricteurs

Dans l'exemple de la figure 3.A, la bascule sur un cycle s'accompagne d'un déphasage entre LI et la mâchoire alors que PL reste phasée avec la mâchoire. L'étude globale du phasage entre la mâchoire et les constricteurs généralise ce résultat. En effet, OC_o advient majoritairement autour de 0% du cycle de mâchoire normalisé (figure 5, droite). En revanche, OL_a est principalement à la fin du geste d'ouverture (cf. /sapa/) ou dans les 3 premiers quart du geste de fermeture (figure 5, à gauche).

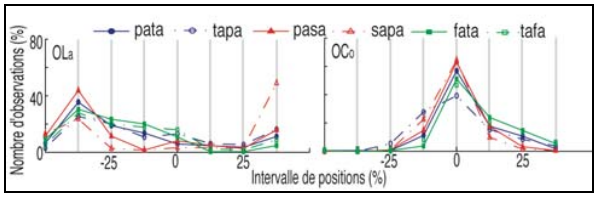


Figure 5 : Position de OL_a et de OC_o sur le cycle de mâchoire (cf. figure 3.C) quand le CVCV se réalise sur un seul cycle.

Phasage entre les constricteurs

Enfin, sur l'exemple de la figure 3.A à gauche, OL_a est plus proche de OC_o que OC_o de OL_{a2}. Ce rapprochement montre que le locuteur évolue vers une structure L_aC_o. L'étude de la durée de OL_a à OC_o par rapport à la durée de OL_a à $OL_{a\,2}$ par locuteur et par séquence (tableau 2) montre que 21/30 moyennes sont inférieures à 50%. Cette dominance se retrouve aussi si on se limite aux moyennes différant significativement de 0 avec un ratio de 16/2. Ainsi, la durée de OL_a à OC_o tend à être plus courte que celle de OC_o à $OL_{a\,2}$.

Tableau 2 : Rapport entre la durée de OL_a à OC_o et la durée de OL_a à OL_a suivant, par séquence et locuteur (en %)

	Pata	Tapa	Pasa	Sapa	Fata	Tafa
S1	49,3	53	50,8	51	53,7**	52
S2	36,3**	40,3*	41,1	43,9	30,5**	36,1**
S3	44,3*	38,6**	41,5**	42,5**	44,8	43,7**
S4	45,4*	49,5	46,3**	52	42,8**	51,4
S5	42,1**	40,8**	51,8	53,5*	37,9**	44,9*

** Moyenne significativement différente de 0, $p < 0.01$ / 30, ttest bilatéral

* Moyenne significativement différente de 0, $p < 0.05$ / 30, ttest bilatéral

Ces résultats confirment la dominance de l'attracteur L_aC_o par rapport à C_oL_a observée dans l'étude acoustique.

6. DISCUSSION ET CONCLUSION

Les résultats des deux études montrent que le système mâchoire-PL-LI est bistable : avec l'accélération, la répétition des CVCV L_aC_o et C_oL_a peut évoluer vers une forme L_aC_o ou C_oL_a. D'autre part, les données articulatoires vont dans le sens du principe d'économie d'énergie et d'une plus grande cohérence articulatoire de L_aC_o par rapport à C_oL_a.

6.1. Multi-déterminisme d'un système bistable

L'évolution vers l'un des attracteurs dépend de différents facteurs. Globalement, la progression vers L_aC_o prévaut, mais est moins nette pour les CVCV C_oL_a que L_aC_o. Ainsi, la coordination de départ et le feed-back auditif pourrait favoriser le maintien de type C_oL_a. D'autre part, la préférence pour L_aC_o est moins marquée pour /pasa/ et /sapa/ que pour les autres couples L_aC_o/C_oL_a ce qui montre un rôle possible du mode d'articulation. Ces deux facteurs pourraient expliquer en partie les profils bistables. Enfin, le fait que certains locuteurs favorisent systématiquement la forme C_oL_a laisse supposer l'implication de variables individuelles telle que la morphologie. Cependant, l'influence de ces facteurs paraît secondaire par rapport à la meilleure cohérence articulatoire L_aC_o puisque le profil L_aC_o domine tant au niveau des locuteurs que des séquences (voir aussi [16]).

6.2. Mâchoire et économie d'énergie

Le deuxième résultat important est que les CVCV L_aC_o et C_oL_a peuvent être produits sur un seul cycle de mâchoire. Ce mode de coordination n'existe pas pour les CVCV dupliqués. Communément, la syllabe CV est associée au cycle de mâchoire. Cette correspondance est au cœur de la théorie FC [5] et de la plupart des études articulatoires. Ainsi, Nelson et al.[14] ont montré que lors de la répétition accélérée d'une syllabe CV les mouvements de la mâchoire progressaient de manière à minimiser la consommation d'énergie. Ils observent aussi une saturation de la durée des mouvements d'ouverture et de fermeture à 50 ms avec une durée préférentielle de 100 ms. Des valeurs similaires sont observées ici. En effet, la durée des cycles de mâchoire (un mouvement de fermeture et un d'ouverture) descend très rarement en deçà de 100 ms. De plus, pour les CVCV L_aC_o et C_oL_a le pic de répartition des durées de cycles est autour de 200 ms. Ainsi, le passage sur un seul cycle permet des durées de CVCV courtes sans saturation de la mâchoire et donc, amoindrit la consommation d'énergie.

6.3. Formes et contraintes substantielles

Ainsi, la production des CVCV avec deux constriction différentes paraît plus économique que celle des CVCV dupliqués. Ce résultat permettrait d'expliquer la dominance des associations syllabiques variant les constrictions dans les lexiques des langues [1][2]. D'autre part, la cohérence articulatoire sous-tendant la forme L_aC_o (avec un geste L_aC_o plus sujet au phasage qu'un geste C_oL_a) permet non seulement d'expliquer la dominance de la forme L_aC_o dans l'étude acoustique mais aussi les asymétries observées dans les transformations verbales [17], en considérant le rôle des interactions perceptuo-motrices dans la perception en général [18] et dans les transformations verbales en particulier [19].

Enfin, replacée dans le courant de recherche visant à dériver la forme de la substance, la stabilité articulatoire des structures L_aC_o constitue un argument pouvant expliquer le fait que les langues préfèrent les structures L_aC_o aux C_oL_a.

BIBLIOGRAPHIE

[1] Rousset, I. "From lexical to syllabic organization: Favored and disfavored co-occurrences", Proc. XVth ICPhS, Barcelona, pp. 715-718, 2003.

[2] Davis, B.L., MacNeilage, P.F. and Matyear, C. "Acquisition of serial complexity in speech", Phonetica, 59(2-3):75-107, Apr-Sep 2002.

[3] MacNeilage, P.F. and Davis, B.L. "On the origins of internal structure of word forms", *Science*, 288:527-531, 2000.

[4] Locke, J. "Movement patterns in spoken language", *Science*, 288:449-451, 2000.

[5] MacNeilage, P.F. "The frame/content theory of evolution of speech production", *Behavioral and Brain Sciences*, 21:499-511,1998.

[6] Vilain, A., Abry, C., Badin, P. and Brosda, S. "From idiosyncratic pure frame to variegated babbling : evidence from articulatory modelling", ICPhS'99, San Fransisco, USA. 1999.

[7] Munhall, K.G., Jones, J.A. (1998), Articulatory evidence for syllabic structure, Behavioural and Brain Sciences, 21 :4, pp.499-521.

[8] Lindblom, B. "On the notion of possible speech sound", *J. of Phonetics*, 18 :135-152, 1990.

[9] Hoyt, D. and Taylor, C.R. "Gait and the Energetic of Locomotion in Horses", *Nature* 292 :239-240, 1981.

[10] Haken, H., Kelso, J.A.S., and Bunz, H. "A theoretical model of phase transitions in human hand movements", *Biological Cybernetics*. 51:347-356, 1985.

[11] Kelso, J.A.S., Saltzman, E.L., and Tuller, B. "The dynamical perspective on speech production: Data and theory", *J. of Phonetics*, 14 :29-59 and 171-196, 1986.

[12] Stetson, R.H. Motor Phonetics: A study of speech movements in action. Amsterdam: North-Holland, 1951.

[13] de Jong, K.J. "Rate-induced resyllabification revisited", *Language and Speech*, 44: 197-216, 2001.

[14] Nelson, W.L., Perkell, J.L. and Westbury, J.R. "Mandible movements during increasingly rapid articulations of single syllables", *J. Acoust Soc Am*, 75(3):945-951, 1984

[15] Sato, M., Schwartz, J.-L,, Abry, C., Cathiard, M.-A. & Loevenbruck, H. (sous presse). Multistable syllables as enacted percept : A source of an assymetric bias in the verbal transformation effect. *Perception & Psychophysics*

[16] Rochet-Capellan, A. and Schwartz, J.-L. "The Labial-Coronal effect and CVCV stability during reiterant speech production: An acoustic analysis", ICPhS, Lisboa, 2005.

[17] Rousset, I., Sato, M., Schwartz, J.-L. and Vallée, N. "Un corrélat perceptif de l'effet LC", *Actes des XXVèmes Journées d'Etudes sur la Parole*, 441-444, 2004.

[18] Schwartz , J.L., Boë, L.J., Vallée, N., and Abry, C. "The dispersion-focalization theory of vowel systems", *J. of Phonetics*, 25 :255-286, 1997

[19] Sato, M., Baciu, M., et al. "Multistable representation of speech forms: An fMRI study of verbal transformations", *NeuroImage*, 23 :1143-1151, 2004.

Stratégie de segmentation prosodique : rôle des proéminences initiales et finales dans l'acquisition d'une langue artificielle

Odile Bagou et Ulrich H. Frauenfelder

Laboratoire de Psycholinguistique Expérimentale, FPSE, 40 bd du Pont D'Arve, 1211 Genève, Suisse
Odile.Bagou@pse.unige.ch

ABSTRACT

Language acquisition in infants and adults depends upon both the segmentation of the words in the speech chain and the extraction of language-specific regularities, particularly prosodic regularities. Two experiments investigate whether prosodic information provided by accented syllables located at the beginning and at the end of prosodic words is used by adult French learners to segment an artificial language. The results allow us to define a prosodic strategy of segmentation : (1) prominence in word final position is used to hypothesize final boundaries ; and (2) cues in word initial position can be used if and only if a primary final prominence is present.

1. INTRODUCTION

Aucun indice acoustique ne marquant systématiquement les frontières lexicales dans la parole continue, la question de savoir quelles informations sont exploitées pour accomplir le processus de segmentation reste entière. Si l'on adopte une approche sous-lexicale de la segmentation, deux principales pistes de recherche ont été explorées dans la littérature. Alors que certains ont mis en évidence le rôle prépondérant des régularités phonotactiques et probabilités transitionnelles dans la segmentation [8][10], d'autres ont suggéré que la segmentation dépendait essentiellement de l'information prosodique [7]. La stratégie de segmentation métrique (MSS) [4] suggère que, dans les langues accentuelles telles que l'anglais, les syllabes fortes indiquent les frontières lexicales initiales potentielles. 87% des mots de contenu étant accentués sur la syllabe initiale en anglais, l'efficacité de cette stratégie est crédible [3]. Cependant, la MSS n'est pas directement applicable à une autre langue. Elle doit s'adapter à la structure métrique de la langue considérée. Le but de cet article est d'expliquer comment l'information prosodique guide la segmentation en français, une langue dans laquelle le domaine de la proéminence est supérieur au mot [5].

Les études récentes sur la parole spontanée admettent que le français possède un système accentuel dual. Ainsi, les accents finaux et initiaux coexisteraient [5]. De par sa nature obligatoire et sa place systématique sur la dernière syllabe pleine des mots de contenu, l'accent final remplirait principalement une fonction de démarcation à droite de l'unité [1]. Frappant les syllabes initiales des mots [5], l'accent secondaire quant à lui, remplirait une fonction démarcative de marquage gauche de l'unité. Ainsi, la coexistence des proéminences initiale et finale servirait à renforcer la cohésion d'une unité et pourrait être utilisée dans la segmentation [6].

Le principal objectif de ce travail est de tester la réalité cognitive de cette fonction démarcative supposée des proéminences initiales et finales dans la segmentation de la parole continue. Plus spécifiquement, puisque l'acquisition requiert une segmentation préalable du signal de parole, nous proposons de préciser la contribution des différentes proéminences, initiale et finale, dans l'acquisition d'une langue artificielle (LA). Enfin, puisque le marquage acoustique des proéminences finales est pluri paramétrique, plusieurs indices pourraient contribuer à la segmentation. L'accent final étant marqué, entre autres, par un allongement final et des variations mélodiques, nous proposons d'explorer la contribution relative de ces indices mélodiques et/ou temporels dans la segmentation de la parole continue.

2. EXPERIENCE 1 : PROEMINENCES FINALES

2.1. Méthode expérimentale

L'usage d'un paradigme d'acquisition de langue artificielle, peu usité en psycholinguistique [1,2,10], nous semblait parfaitement répondre à nos attentes : d'une part, parce qu'il permet de mettre des adultes disposant de connaissances linguistiques préalables dans une situation d'acquisition et d'autre part, parce qu'il permet la manipulation des facteurs d'intérêt.

Mode de passation et tâche

La passation se déroule en deux temps : (1) Une phase d'apprentissage et (2) une phase test, d'évaluation de la performance des participants.

Lors de la première étape, la langue (e.g. /pyk.tug.bõf.vob.daz.nul.chije.syn.tug.dõʀ.kãv.nøn…/) est soumise auditivement aux participants. Pendant ce bain de langage, la tâche est de retenir les séquences de syllabes qui constituent un "mot" de la langue. Aucune information relative à la taille et à la structure des unités n'est donnée.

Lors de la seconde étape, une tâche de préférence lexicale à choix forcé sans contrainte temporelle réelle permet de vérifier l'intégration mnémonique des unités artificielles. Les participants sont soumis à des paires présentées auditivement, constituées de deux extraits de la langue correspondant soit à un "mot" et un "non-mot", soit à deux "non-mots". La tâche est de choisir celui des deux stimuli qui correspond ou ressemble plus à une "unité lexicale" de la langue artificielle apprise préalablement.

Construction de la langue de base

La langue artificielle de base est le résultat de la concaténation de 18 syllabes respectant les règles phonotactiques du français. Enregistrées par un locuteur francophone suisse romand, les syllabes sont ensuite re-synthétisées selon la méthode PSOLA. Ainsi, les variations de fréquence fondamentale ont été harmonisées sur chacune des syllabes pour obtenir une fréquence uniforme de 110 Hz, correspondant au fondamental moyen du locuteur. Huit mots artificiels bi- et trisyllabiques ont alors été créés par concaténation. Enfin, cent occurrences de chacun des huit mots artificiels ont été concaténées dans un ordre semi-aléatoire en une séquence de parole continue, sans pause ni indice signalant la présence d'une frontière. Le même ordre de présentation était utilisé dans les différentes versions de la LA.

Conditions expérimentales

Quatre versions de la LA ont été construites : une LA de base (A) sans indices prosodiques de frontière : les auditeurs pouvaient inférer les unités en calculant des statistiques de co-occurrence des syllabes adjacentes (Probabilités transitionnelles PT) ; et 3 LA dont les unités portaient une proéminence finale (B, C, D). Dans ces versions prosodiquement marquées, au moins un indice acoustique caractéristique de la proéminence finale était fourni en plus de l'information statistique. Les indices mélodiques et temporels étaient tous deux disponibles dans la version D, alors que dans les versions B et C, seul l'un des deux indices était fourni : l'indice d'allongement (B) ou l'indice mélodique (C).
Les manipulations acoustiques ont été réalisées par le logiciel Praat, lequel nous a permis d'allonger la dernière syllabe de chaque unité de 30% de sa durée intrinsèque (B et D) et/ou d'augmenter la hauteur mélodique de la dernière syllabe de 110 à 130 Hz (C et D).

Construction du test

Chaque test était constitué de 80 paires tests : 32 paires impliquant un mot (M) et un non mot (NM) et 48 paires impliquant deux non mots. Les paires tests étaient construites en appariant chacun des 8 mots artificiels avec les 32 non mots partageant certaines caractéristiques des mots. Seules les analyses menées sur les paires M-NM sont présentées dans cet article.

Participants

96 Suisses romands ont été répartis en 8 groupes expérimentaux de 12 participants chacun : 4 groupes tests appariés à 4 groupes contrôles, respectivement exposés ou non à la langue artificielle durant la phase d'apprentissage. De plus, l'ordre de présentation des paires testées était contrebalancé pour annuler des artefacts potentiels d'apprentissage par la tâche. Ainsi, la moitié des participants d'un groupe répondait à la première partie du test avant la deuxième, et inversement pour l'autre moitié de participants.

2.2. Résultats

Des t-tests ont été calculés afin de savoir si les participants obtenaient une performance supérieure au seuil de hasard (50%) dans chaque condition. Tous sont significatifs sauf pour le groupe contrôle *($t(47)=-.25$, $p=.8$)*, ce que nous attendions puisque ces participants n'avaient pas subi la phase d'apprentissage préalable.
Des ANOVAs à mesures répétées ont été réalisées - l'une considérant les participants (F_1) et l'autre, les items (F_2) comme variable aléatoire - avec le taux de réponses correctes comme variable dépendante. Les analyses contenaient un facteur intra-sujets, le type de paire, et trois facteurs inter-sujets à savoir, la condition expérimentale, l'ordre de présentation des blocs dans le test et le groupe expérimental. Les résultats sont présentés dans la Table 1.

Table 1 : Taux de réponses correctes moyen (%) selon de groupe et les indices disponibles lors de l'acquisition

Versions LA	**Indices disponibles lors de l'acquisition**			
	A (PT)	B (PT + All)	C (PT + f0)	D (PT+ All. +f0)
Test	64.6	82.0	89.8	85.4
Contrôle	51.8	47.9	51.3	47.7
Δ	**12.8***	**34.1****	**38.5****	**37.7****

La significativité de l'effet du groupe (contrôle vs. test) suggère que la phase d'apprentissage est nécessaire à l'acquisition et à la rétention des mots artificiels *($F_1(1,80)= 274.2$, $p<.0001$; $F_2(1,6)=112.2$, $p<.0001$)*. Une amélioration de la performance (de 30% en moyenne) due à l'apprentissage préalable apparaît et ce, quelle que soit la condition expérimentale, i.e. quels que soient les indices disponibles dans la langue apprise.
De plus, l'effet significatif de la condition expérimentale suggère que les participants soumis à une langue dont les frontières finales d'unité sont indicées par des informations prosodiques obtiennent de meilleures performances que ceux qui ne disposent que d'indices statistiques de co-occurrence des syllabes pour intégrer les mots *($F_1(3,80)=7.7$, $p=.0001$; $F_2(3,18)=5$, $p=.01$)*. Ainsi, ces résultats indiquent que l'allongement final et/ou les variations mélodiques permettent une segmentation plus précise de la chaîne de parole que le calcul statistique seul.
Les comparaisons multiples (test de Tukey) indiquent également que la différence d'environ 6% observée entre les performances des participants soumis à la condition C (f0) et celles des participants soumis à la condition B (all.) tend à la significativité *($p=.1$)*. Ainsi, les variations mélodiques pourraient constituer un indice de frontière plus puissant que l'allongement final.
Par ailleurs, le cumul d'informations acoustiques (D : f0+allongement final) n'améliore pas significativement la performance. En effet, aucune différence n'est avérée entre les conditions dont les mots ne portent qu'un seul indice acoustique d'allongement (B) ou de f0 (C) et la condition D (respectivement, *$p = .9$; $p = .4$*).
Enfin, l'effet de la condition expérimentale interagit significativement avec le groupe expérimental *($F_1(3,80)=10.7$, $p<.0001$; $F_2(3,18)=7.8$, $p=.002$)*. Ainsi, la différence de performance entre les groupes contrôle et

test est plus conséquente lorsque les indices prosodiques de frontière sont présents que lorsque seules les informations statistiques sont disponibles (37 % en moyenne contre 12.8% dans la condition sans indice prosodique).
En résumé, qu'elle soit marquée par un allongement ou des variations mélodiques, la présence d'une proéminence finale permet un apprentissage plus efficace. Cependant, la redondance du marquage n'apporte aucune information supplémentaire.

3. EXPERIENCE 2 : PROEMINENCES INITIALES

3.1. Méthode expérimentale

La méthode générale de passation est identique à celle de l'expérience précédente (voir chap.2.1.). Seules les caractéristiques de la langue diffèrent.

Construction de la langue de base

La langue artificielle de base est constituée de 24 syllabes respectant les règles phonotactiques du français. Enregistrées par une locutrice francophone suisse romande, les syllabes sont ensuite re-synthétisées grâce au logiciel Praat selon la méthode PSOLA. Ainsi, les variations de fréquence fondamentale sont harmonisées pour obtenir une fréquence uniforme de 210 Hz, correspondant au fondamental moyen de la locutrice. Huit mots artificiels trisyllabiques, ont été créés par concaténation. Enfin, 96 occurrences de chacun des huit mots ont été concaténées dans un ordre semi-aléatoire en une séquence de parole continue, sans pause ni indice signalant la présence d'une frontière.

Conditions expérimentales

Trois versions de la LA ont été générées : une LA de base sans indices prosodiques de frontière (« plate ») : les frontières d'unités pouvaient être inférées en considérant que des probabilités transitionnelles faibles entre les syllabes signalent la présence d'une frontière. Par ailleurs, 2 LA dont les unités portaient une proéminence initiale accompagnée ou non d'une proéminence finale ont été construites. Dans la version « Ai », seule la proéminence initiale était présente et marquée par une fréquence fondamentale variant de 230Hz au début de syllabe, à 250Hz au centre du noyau vocalique [11]. Dans la version « Arc » [6], les unités portaient des proéminences initiales et finales. Pour le marquage de la proéminence finale, la fréquence fondamentale moyenne de 210Hz était élevée à 250Hz au début de la syllabe pour atteindre 270 Hz, au centre du noyau vocalique. De plus, la dernière syllabe de chaque unité était allongée de 30% de sa durée intrinsèque. Pour le marquage de la proéminence initiale, les caractéristiques étaient similaires à celles de la langue « Ai ». Les manipulations acoustiques ont été réalisées grâce au logiciel Praat.

Participants

51 francophones monolingues, étudiants en psychologie à l'université de Genève, ont été répartis en 3 groupes expérimentaux de 17 participants chacun.

3.2. Résultats

Des t-tests ont été calculés afin de savoir si les participants obtenaient une performance supérieure au seuil de hasard dans chaque condition. Tous sont significatifs, ce qui suggère que les participants n'ont pas répondu au hasard (50%).
Des ANOVAs à mesures répétées ont été réalisées - l'une considérant les participants (F_1) et l'autre, les items (F_2) comme variable aléatoire - avec le taux de réponses correctes comme variable dépendante. Les analyses contenaient un facteur intra-sujets, le type de paire, et un facteur inter-sujets, la condition expérimentale. Un effet principal de la condition expérimentale significatif apparait ($F_1(2,48)=6.2$, $p=.004$; $F_2(2,14)=9.5$, $p=.002$) (Figure 1).

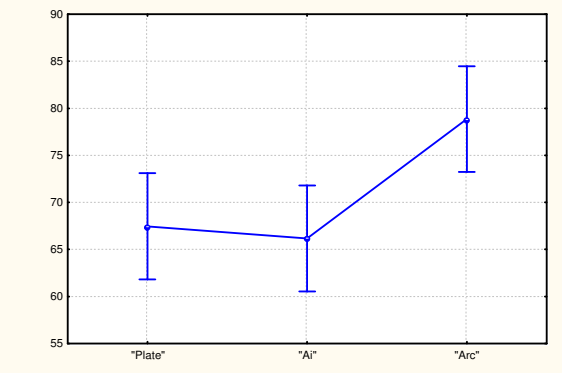


Figure 1 : Taux de réponses correctes moyens (%) selon le type de proéminence disponible durant l'apprentissage

Les comparaisons multiples (test de Tukey) confirment que la différence de performance entre les versions « plate » et « Arc » de la langue est significative ($p=.02$). Ainsi, le bornage bipolaire des unités artificielles fournit un indice pertinent de segmentation et favorise l'acquisition des unités. De même, la différence de performance entre les versions « Ai » et « Arc » est significative ($p=.007$), ce qui suggère que la proéminence initiale seule n'est pas aussi efficace que l'arc accentuel.
Par ailleurs, la performance n'est pas améliorée par la présence unique d'une proéminence initiale (« plate » vs. « Ai » : $p=.9$). Ainsi, l'accent initial ne permettrait pas de segmenter plus efficacement le signal de parole.

4. DISCUSSION GENERALE

Les résultats des deux expériences indiquent que l'information portée par les syllabes accentuées facilite l'acquisition d'une nouvelle langue. Dans les versions A (exp. 1) et « plate » (exp. 2), les participants ne disposaient que de l'information phonotactique et distributionnelle (PT). Leur faible performance corrobore le fait que ces indices ne sont pas suffisants pour segmenter idéalement le signal de parole. En revanche, la présence d'informations prosodiques sur les syllabes finales facilite l'acquisition de

la langue : l'expérience 1 atteste de la pertinence des proéminences finales dans l'acquisition et l'expérience 2 indique que le bornage bipolaire des unités facilite leur intégration mnémonique.

L'expérience 1 a été également menée pour explorer la contribution relative de différents indices typiques des proéminences finales. Puisque les participants de la condition B obtiennent des performances comparables à ceux de Banel et al. [2], l'effet de l'allongement final a été répliqué. Etonnement, les variations mélodiques finales seules induisent des réponses plus précises que l'allongement final seul. Ce résultat infirme les conclusions de Rietveld [9], lequel suggère que l'allongement final est l'indice le plus efficace pour segmenter des phrases dont la frontière est ambiguë. Cependant, dans notre LA, l'occurrence régulière d'un contour continuatif toutes les 2 ou 3 syllabes a pu faciliter la segmentation puisque les auditeurs pouvaient supposer qu'ils écoutaient une liste d'items plutôt que des phrases. Par ailleurs, on pourrait supposer que cette efficacité de la mélodie résulte de l'impression subjective d'allongement qu'elle induit. Cependant, si la bonne performance était simplement due à un cumul « virtuel » d'indices, alors elle aurait dû être améliorée lorsque les indices étaient effectivement cumulés (version D). Ainsi, cette prééminence des variations mélodiques mérite d'être répliquée.

Alors que le bornage bipolaire (« arc ») permet une meilleure rétention des unités artificielles, la présence unique de la proéminence initiale (« Ai ») n'améliore pas l'acquisition. Puisqu'en français, les constituants prosodiques ont la « tête » à droite, il était raisonnable de penser que les participants auraient attribué un statut de proéminence primaire finale à cette syllabe saillante. Ainsi, les frontières signalées par l'information prosodique auraient été différentes de celles suggérées par le calcul statistique (PT). Le fait que la performance dans la version « Ai » soit comparable à celle de la version « plate » aurait donc suggéré que les participants avaient privilégié les frontières indicées par les PT. Toutefois, nous avons montré que l'acquisition était facilitée par la présence d'une proéminence finale (exp.1), ce qui nous invite à rejeter l'idée que cette syllabe saillante ait été interprétée comme une proéminence finale.

De plus, puisque les patrons appliqués à nos unités artificielles ont été observés en début de mot de contenu dans des productions naturelles [11], il est probable qu'ils aient été interprétés comme des contours typiques de la proéminence initiale dans notre langue artificielle. En effet, le système pourrait être capable de détecter des motifs prosodiques fréquemment rencontrés aux frontières d'unités, ce qui lui éviterait de confondre des proéminences initiales et finales. Dans ce cas, nos résultats indiquent que la proéminence initiale n'a pas été utilisée pour segmenter les unités artificielles, ce qui suggère que les proéminences initiales ne rempliraient peut être pas le rôle démarcatif qu'on leur attribue communément.

Cependant, cette absence d'effet pourrait être due à la spécificité du matériel. Puisque la longueur syllabique de toutes les unités artificielles était identique, les participants ont pu appliquer une stratégie triviale de dénombrement de syllabes, plutôt que d'utiliser l'information prosodique pour découvrir les mots de la langue. Cette explication est toutefois peu probable puisque nous observons, avec un matériel comparable, une amélioration nette de la performance due à la présence conjointe de proéminences initiales et finales (« Arc »).

Par ailleurs, il est probable que le suisse romand soit marqué par ses origines franco-provençales dont la structure prosodique est fortement paroxytonique. Ainsi, les caractéristiques que nous avons données à la montée initiale pourraient ne pas être celles que les participants attendaient.

Enfin, les unités de notre LA sont des « pseudo-mots de contenu » qu'aucun « pseudo-mot grammatical » ne sépare. Or Welby [11] montre que la montée mélodique initiale caractérise une frontière entre mot grammatical et mot de contenu. Ainsi, la montée mélodique n'aurait pu remplir sa véritable fonction dans notre langue, ce qui explique qu'elle n'ait pu intervenir dans l'acquisition.

Pour conclure, la stratégie de segmentation prosodique serait principalement basée sur les proéminences finales dont la présence constituerait une condition nécessaire à l'usage éventuel des proéminences initiales. Reste à préciser si leur fonction est de signaler la fin de l'unité en cours de traitement ou, plus plausiblement, d'attirer l'attention vers le début de l'unité suivante.

5- BIBLIOGRAPHIE

[1] Bagou, O., Fougeron, C., Frauenfelder, U. H. Contribution of Prosody to the Segmentation and Storage of "Words" in the Acquisition of a New Mini-Language, Bernard Bel & Isabelle Marlien (eds.), *Proceedings of the Speech Prosody 2002 conference,* 11-13 April 2002, Aix, France, 59-62, 2002.

[2] Banel, M. H., Frauenfelder, U. H. et Perruchet, P. Contribution des indices métriques à l'apprentissage d'un langage artificiel. *JEP. Martigny*, Suisse, 29-32, 1998.

[3] Cutler, A. et Carter, D. M. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142, 1987.

[4] Cutler, A., Norris, D. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology : Human Perception and Performance*, 14, 113-121, 1988.

[5] Di Cristo, A. Vers une modélisation de l'accentuation du français : seconde partie, *French Language Studies*, 9, 143-179, 2000.

[6] Fonagy, I. L'accent français : accent probabilitaire, *Studia Phonetica*, 15, 123-133, 1979.

[7] Mattys, S. L., Jusczyk, P. W., Luce, P. A., Morgan, J. L. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38(4), 465-494, 1999.

[8] McQueen, J. M. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language,* 39, 21-46, 1998.

[9] Rietveld, A. C. M. French word boundaries, *Language and Speech*, 23(3), 289-296, 1980.

[10] Saffran, J. R., Newport, E. L., Aslin, R. N. Word segmentation : The role of distributional cues. *Journal of Memory and Language*, 35(4), 606-621, 1996.

[11] Welby, P. The Slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation, thèse de Doctorat, Université Ohio, 2003.

Tomber le masque de l'information: effet *cocktail party*, masque informationnel et interférences psycholinguistiques en situation de compréhension de la parole dans la parole.

Michel Hoen[1,2], Claire-Léonie Grataloup[1], Nicolas Grimault[2], Fabien Perrin[2], Xavier Perrot[2], François Pellegrino[1], Fanny Meunier[1], Lionel Collet[2]

[1]Laboratoire dynamique du Langage
UMR5096 CNRS, Université Lumière, Lyon, France

[2]Laboratoire Neurosciences et Systèmes Sensoriels
UMR5020 CNRS, Université Claude Bernard, Lyon, France

michel.hoen@phonak.ch

ABSTRACT

Up to now, the comprehension of speech in noise and more particularly in concurrent speech sounds was rarely studied in the domain of psycholinguistics. In this paper we report a study testing the differential effects of speech derived noises as multi-talker cocktail party sounds and their time-reversed pendent on the comprehension of isolated words. Results suggest that different levels of linguistic information from concurrent speech signals can compete with linguistic information in target signals, mainly depending on the spectral saturation caused by increasing the number of voices in concurrent signals. These results suggest linguistically specific participations in informational masking effects occurring in the context of speech in speech comprehension.

1. INTRODUCTION

Bien que nous soyons aptes à comprendre de la parole distillée par les casques d'une chambre anéchoique, nous sommes plus souvent confrontés à la situation où la parole nous parvient du chaos acoustique d'un grand carrefour ou de l'ambiance criarde d'une salle de réunion. Pourtant, nous restons souvent capables d'en comprendre le message. La parole reste intelligible dans des conditions acoustiques extrêmement variables et malgré la présence de quantités importantes de bruits interférents. Ce phénomène, appelé « *Effet Cocktail Party* » correspond à une faculté cognitive spécialisée nous permettant de focaliser notre attention sur un flux auditif particulier parmi différents flux concurrents. Depuis sa première description par Cherry en 1953 [1], l'effet cocktail party a donné lieu à un grand nombre d'études ayant permis de cerner certains de ses fondements cognitifs. Il a été par exemple démontré que l'effet cocktail party reposait sur la capacité du système auditif humain à réaliser une séparation spatiale des flux concurrents [2]. Cependant, lorsque les flux acoustiques ont une origine spatiale commune, ou lorsque le son est présenté de façon diotique (le même signal aux deux oreilles), le système cognitif doit alors pouvoir utiliser d'autres indices.

Dans ce contexte, différentes études ont pu mettre en évidence l'importance d'indices temporels lents ou d'indices de surface, comme des différences globales de fréquence fondamentale (F_0), d'accent, de style discursif ou encore d'intensité entre les flux à séparer [3-8]. En situation de compétition de flux acoustiques, on décrit souvent les effets de masques attribuables aux flux interférents comme pouvant agir à deux niveaux principaux : un niveau énergétique et un niveau informationnel [7-8]. L'effet de masque énergétique est dû aux propriétés spectrotemporelles des sons concurrents. Il se produit lorsque la parole est émise en présence d'un bruit à large bande spectrale qui va se superposer partiellement, en temps ou en fréquence, au signal de parole cible. L'effet de masque informationnel est quant à lui attribuable au type d'information contenue dans le bruit [4-5]. Dans ce cas, ce sont les informations que les signaux véhiculent qui vont entrer en compétition et perturber l'interprétation cognitive qui sera faite du signal cible. Le phénomène de masque informationnel est particulièrement saillant lorsqu'un signal de parole est émis en présence d'autres signaux de parole. Dans ce contexte particulier, Brungart et coll., ont étudié l'intelligibilité d'un signal de parole cible en fonction du nombre de voix concurrentes (2, 3 ou 4 locuteurs) et du Rapport du Signal au Bruit (RSB) de signaux de parole interférents [7-8]. Leurs résultats ont montré une décroissance linéaire des performances en fonction du RSB, dès que trois voix au moins étaient mises en concurrence. La condition à deux voix concurrentes étant résolue sur la base d'indices de surface permettant de discriminer les voix (modulations de F_0, timbre ou style discursif). Ces études montrent la disparition de l'impact de tels indices acoustiques de surface dans la résolution de l'effet cocktail party à partir de situations d'interférence à trois voix de genre identique (ou à partir de 4 voix concurrentes de genre mixte). Ainsi, il semble qu'au-dessous de quatre voix concurrentes, la résolution de l'effet cocktail party repose essentiellement sur l'utilisation d'une stratégie de séparation spatiale des flux ou d'indices acoustiques de surface. Bien que la caractérisation de l'effet cocktail party ait donné lieu à un très grand nombre d'études, peu d'expériences se sont intéres-

sées aux interférences fines ayant lieu au-delà de 4 voix concurrentes. En particulier, les effets de masques psycholinguistiques pouvant apparaître dans les situations de compréhension de la parole dans la parole n'ont jamais été mis en évidence. Pourtant, le concept de masque informationnel prend dans ce contexte un sens particulier puisqu'il pourrait alors être attribué aux différents types d'information linguistique contenue dans la parole, telles que l'information prosodique, phonologique ou lexicale. On pourrait alors aisément imaginer, au sein du concept de masque informationnel, l'existence de sous-types d'effets liés à l'accessibilité dans le son concurrent de ces différents niveaux d'information linguistique. Cette accessibilité devrait dépendre de la transparence spectro-temporelle du signal interférent, dès lors, la sensibilité des sujets à ces différents niveaux d'information serait hypothétiquement modulée par la saturation énergétique du signal (proportionnelle au nombre de locuteurs) ainsi que par le RSB.

La présente expérience a été conduite afin de mettre en évidence, au sein du masque informationnel, l'existence de différents types d'interférences linguistiques dans des situations de compréhension de la parole dans la parole. Nous avons pour cela conduit une expérience de compréhension de mots isolés en présence de bruits paroliers. Dans ces bruits, nous avons manipulé le nombre de locuteurs et la nature physique exacte des signaux. Nous avons comparé les effets de masques dus à des enregistrements de 'cocktail party' standards à 4, 6 ou 8 voix simultanées et ceux dus aux mêmes enregistrements mais inversés selon leur dimension temporelle. L'inversion temporelle de la parole, ou 'reversed speech', a parfois été considérée comme la manipulation la plus drastique pouvant être appliquée à un signal de parole [9]. En réalité, la parole inversée conserve les propriétés énergétiques du signal de parole source, mais 'sonne' aussi comme la parole naturelle, puisqu'elle en conserve certains traits infra-segmentaux. Les voyelles en particulier sont bien conservées et certaines consonnes présentent de bons degrés de réversibilité. Mieux encore, lorsque plusieurs flux de parole inversée sont superposés, le signal composite sonne comme un signal de cocktail party, et des phonèmes peuvent y être perçus. En revanche, l'information lexicale est totalement perdue. Afin de pouvoir différencier les effets de masque linguistiques du simple effet de masque énergétique, nous avons également employé comme son masquant un bruit large bande ayant les mêmes propriétés spectrotemporelles qu'un bruit de cocktail party mais ne contenant pas d'information linguistique.

2. MATERIELS ET METHODES

2.1. Bruits Interférents

Bruits de *cocktail party* multi-locuteurs

Trois bruits de cocktail party ont été créés en mixant 4, 6 ou 8 enregistrements de locuteurs uniques. Chaque locuteur a été enregistré individuellement à l'aide d'une chaîne d'acquisition et de numérisation comportant un microphone Røde NT1, un préamplificateur Ultragain MIC 2000 et une carte son RolandED UA-30, les sons étant numérisés à 44 kHz sur 16 bits. Chaque source individuelle consistait en une voix masculine ou féminine prononçant des phrases intelligibles en langue française, dont les noms propres avaient été supprimés. Les voix sources ont toutes subi la même chaîne de traitement: i) suppression des pauses et silences excédant une seconde; ii) suppression des portions d'enregistrements contenant des erreurs de prononciation ou des marques prosodiques inappropriées; iii) réduction du bruit de fond optimisée pour les signaux de parole (CoolEdit Pro© 1.1 – Dynamics Range Processing – preset Vocal limiter); iv) calibration en dBA et normalisation de chaque source à 80 dBA (Larson Davis System LD824 et oreille artificielle: AEC101); et enfin : v) mixage des différentes sources et sauvegarde au format .wav (44kHz, 16 bits, Stéréo).

Bruits de *cocktail party* inversés

Les bruits de cocktail party inversés ont été obtenus en appliquant une inversion point à point selon l'axe temporel des bruits de cocktails multi-locuteurs décrits ci-dessus. Ceci fut réalisé grâce à une routine implémentée dans le logiciel Adobe® Audition™ dans sa version 1.0.

Bruit large bande associé

Afin d'obtenir un bruit à large spectre aux propriétés énergétiques semblables à celles de nos bruits paroliers nous avons décidé de partir du son cocktail party comprenant 8 locuteurs (celui-ci ayant le spectre énergétique le plus large et le plus dense) et d'en dériver un bruit ne contenant plus aucune information linguistique. Pour ce faire, nous avons commencé par extraire l'enveloppe temporelle du bruit original sous 60Hz, afin d'en dériver les fluctuations dynamiques lentes. Puis, par une transformée de Fourier (FFT), nous avons calculé l'énergie spectrale du signal d'origine et en avons extrait la distribution des phases. Les phases ont été redistribuées de façon aléatoire, puis réinjectées dans l'enveloppe temporelle du bruit de cocktail party original. Enfin, l'énergie globale rms du bruit obtenu a été ajustée à celle du signal original. Le bruit résultant possède la même énergie spectrale et la même enveloppe que le bruit original, mais les phases étant aléatoires, il ne comporte plus aucune information d'ordre linguistique.

2.2. Mots Cibles

320 mots français monosyllabiques, tri-phonémiques ont été enregistrés. Ils ont été sélectionnés dans une gamme de fréquence d'occurrence moyenne (0.19 occurrences par million (opm) à 146.71 opm ; moy = 20.96 ; DS = 21.37) d'après Lexique2 [10], ceci afin d'éviter des items de trop haute ou trop basse fréquence. Les mots isolés étaient prononcés par un locuteur masculin unique et enregistrés en chambre sourde à l'aide d'un microphone Sony ECM-MS907 et sauvegardés au format .wav (44 kHz, stéréo, 16 bits).

2.3 Conditions et Sujets

L'effet de masque de 7 types de bruits sur la compréhension de mots cibles isolés a été testé: trois bruits de cocktail party à 4, 6 et 8 voix mixtes, trois bruits de cocktails inversés à 4, 6 et 8 voix mixtes et un bruit à large bande. Chacun de ces bruits était testé pour des RSBs de -3, 0, +3 et +6 dBs, générant un total de 28 conditions. 36 sujets ont pris part à l'expérimentation. Nous avons généré 36 listes de 8 mots cibles, équilibrées en fréquence et en nombre de voisins phonologiques. Les mots dans les listes finales avaient une fréquence de 3.92 opm (DS = 0.006) et un nombre de voisins phonologiques de 19.83 (DS = 0.06). Les participants étaient tous étudiants, âgés de 18 à 32 ans, de langue maternelle française et dépourvus de déficits auditifs ou langagiers diagnostiqués, ils étaient rémunérés pour leur participation.

2.4 Stimuli

Les stimuli étaient 288 fichiers au format .wav (44 kHz, 16bits, stéréo) ayant chacun une durée de 4 secondes. Le bruit de fond était présent durant toute la durée du stimulus et le mot cible était systématiquement inséré à 2.5s du début du fichier. Les mots cibles avaient des durées variables de 242.68 ms à 914.47 ms (moy = 550.32 ms, DS = 134.56 ms). Les extraits de bruits ont été sélectionnés au hasard et contrebalancés. Par ailleurs, comme le mixage final du bruit et des mots cibles résultait en des variations d'intensité globale des stimuli, nous avons appliqué suite au mixage, une normalisation en intensité aléatoire sur une gamme de $^{+}/_{-}$ 3 dB par pas de 1dB. L'intensité globale des stimuli obtenus ne pouvait ainsi être prédictive de la condition de stimulation.

2.5 Procédure

Les participants étaient assis face à un écran d'ordinateur. Les stimuli étaient présentés de façon diotique au moyen d'un casque audio (Beyerdynamic DT 48, 200Ω) à un niveau d'écoute confortable fixé individuellement. Tous les bruits, listes de mots et conditions étaient aléatoirisées parmi les sujets. La tâche consistait à écouter les stimuli et à retranscrire au clavier le mot cible ou bien la portion de mot cible entendu. Les sujets étaient familiarisés au décours temporel des essais sur 12 exemples. La durée totale de l'expérience allait de 30 à 60 min en fonction de l'habileté des sujets à utiliser un clavier informatique. Les transcriptions des sujets étaient enfin analysées en termes de proportion de mots correctement reproduits.

3. RESULTATS

Afin de tester l'effet du nombre de voix présentes dans un bruit de cocktail party interférent sur la compréhension d'un signal de parole cible, nous avons réalisé une première Analyse de variance (Anova) en prenant les taux de récupération lexicale individuels comme variable dépendante.

L'analyse incluait les facteurs de bruit: 'Type' (2: Cocktail Party ou Cocktail Inversé), 'Nombre' de voix présentes dans le bruit (3: 4, 6 ou 8) et le RSB (4 : -3, 0, +3 et +6dB), (voir Fig. 1). Cette analyse a révélé un effet principal uniquement pour le facteur RSB ($F(3,105) = 185,31$; $p<.05$), le taux de récupération lexicale décroissant de façon linéaire avec le RSB dans notre gamme de RSBs testés. Les deux autres facteurs, 'Type' et 'Nombre' n'avaient pas d'effets principaux significatifs, respectivement $F(1,35) = 1,18$; n.s. et $F(2,70) = 2,19$; n.s., suggérant une absence d'effet global de ces facteurs. En revanche, l'interaction de second ordre entre ces deux facteurs était significative ($F(2,70) = 4,53$; $p<.05$). Aucune autre interaction n'était significative, en particulier, le facteur RSB n'interagissait ni avec le facteur 'Type' ($F(3,105) = 0,93$; n.s.), ni avec le facteur Nombre ($F(6,210) = 0,50$; n.s.), suggérant une décroissance globale des performances avec le RSB indépendante de la condition de bruit considérée. Le facteur RSB a pour cette raison été ensuite supprimé des autres analyses en moyennant l'ensemble des données obtenues aux mêmes RSBs. Un test post-hoc de type LSD ($\alpha=.05$), a ensuite été appliqué à l'interaction de second ordre significative afin d'établir l'influence du nombre de locuteurs sur la façon dont les différents types de bruits masquaient les mots cibles. Au sein des bruits paroliers de type cocktail party, cette analyse révélait un effet non linéaire du nombre de locuteurs puisque le bruit à 6 voix avait l'effet de masque le moins important et différait significativement des effets de masques des bruits de cocktail à 4 voix ($p=0.004$), et à 8 voix ($p=0.056$). Les deux bruits de cocktails à 4 et 8 voix avaient les effets de masque les plus importants. En comparant les effets de masques dus aux bruits de type cocktail party à ceux dus aux bruits de type cocktail inversé, une différence significative était observée entre le cocktail à 4 voix et le cocktail inversé à 4 voix ($p=0.004$), le bruit de cocktail à 4 voix étant associé à l'effet de masque le plus important. Bien que l'effet de masque dû aux bruits de cocktails inversés semble décroître avec le nombre de voix impliquées, cette décroissance est trop lente pour aboutir à des différences significatives entre bruits de cocktail inversés à 4, 6 ou 8 voix, nous considérerons donc que ces trois bruits ont des effets de masques très comparables.

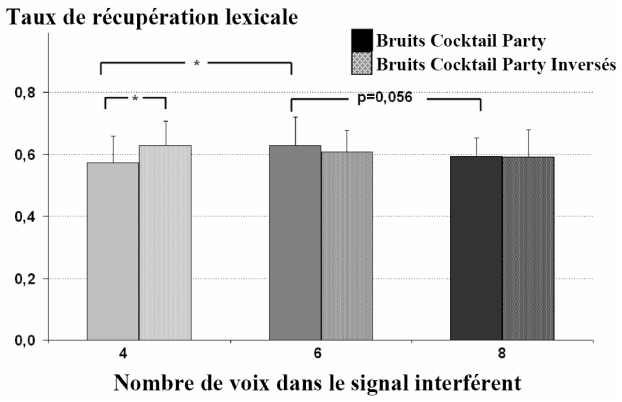


Figure 1: *Taux de récupération de mots isolés en fonction du type de bruit parolier interférent.*

Dans l'analyse précédente nous avons observé une différence significative entre les effets de masques dus au bruit cocktail party à 4 voix et au même bruit inversé. Cette observation suggèrerait l'existence de deux niveaux de masquage informationnel dans une condition à 4 locuteurs, uniquement si ces deux effets sont eux-mêmes distincts d'un effet de masque purement énergétique.

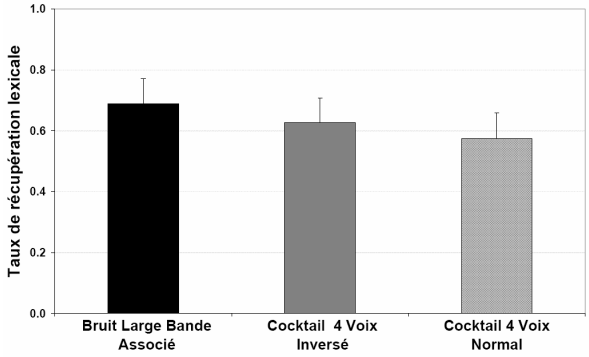


Figure 2: *Taux de récupération de mots isolés pour trois types de bruits concurrents: le bruit associé, le cocktail party inversé à 4 voix et le cocktail party à 4 voix.* ***(Toutes différences significatives)***

Afin d'élucider cette question nous avons réalisé une seconde Anova à un niveau, en prenant la moyenne des taux de récupération lexicale obtenus au même RSB comme variable dépendante et pour seul facteur le 'Type' de bruit (3: Bruit associé, Cocktail Party 4 voix inversé et Cocktail Party 4 voix). Cette analyse révélait un effet principal significatif du 'Type' de bruit considéré ($F(2,70)=23,45$; $p<.05$) et les comparaisons planifiées étaient toutes significatives, montrant que ces trois types de bruits avaient des effets de masque significativement différents. Le bruit large bande associé avait le moins grand effet de masque, puis venait le bruit de cocktail party à 4 voix inversé et enfin le bruit de cocktail party à quatre voix (voir Fig. 2).

4. DISCUSSION

Les résultats de cette expérience sur l'effet de masque informationnel en situation de cocktail party multi-locuteur ont permis de mettre en évidence, au sein de l'effet de masquage informationnel, de potentiels effets de l'accessibilité de différentes sources d'information linguistique dans un bruit parolier concurrent. En effet, nous avons pu montrer qu'un bruit physique associé, un bruit de cocktail inversé et un bruit de cocktail party standard ayant tous trois des propriétés énergétiques comparables avaient trois effets de masquage distincts, le signal de cocktail party normal ayant l'effet de masquage le plus fort. Ces trois niveaux de masque pourraient être attribués respectivement à un niveau énergétique pur (bruit large bande associé), à l'ajout d'un niveau d'ordre phonologique (bruit de cocktail inversé) et enfin à un effet combiné énergétique, phonologique et lexical dans le cas du bruit de cocktail party à 4 voix standard. L'accessibilité d'information lexicale dans le cocktail à 4 voix étant attestée par certaines erreurs des sujets ayant répondu, rarement mais de façon symptomatique, avec des mots issus du bruit de cocktail plutôt qu'avec les mots cibles. Cet effet de masque lexical disparaît dans notre expérience pour des nombres de locuteurs supérieurs à 4 voix, le cocktail party à 6 locuteurs étant même le bruit de cocktail le moins masquant. Ceci pourrait être dû à une disparition de l'accessibilité de l'information lexicale dans le bruit de cocktail party du fait d'une progressive saturation spectrale du signal causée par l'ajout progressif de locuteurs. Ce résultat inédit offre un nouveau cadre d'étude pour les mécanismes de compétition d'informations ayant lieu lors de l'accès au lexique mental puisqu'il met pour la première fois en évidence des compétitions d'informations psycholinguistiques dues au contenu informationnel d'un signal interférent. Ce paradigme pourrait permettre de tester directement des hypothèses sur les phénomènes de compétition d'information lexicale dans le cadre de l'accès au lexique telles que décrites dans les modèles psycholinguistiques.

REMERCIEMENTS

Cette étude a été réalisée grâce aux fonds de l'ACI (n° 67068) du *Ministère de l'Enseignement Supérieur et de la Recherche français,* attribuée à Fanny Meunier.

BIBLIOGRAPHIE

[1] Cherry, E. (1953). ''Some experiments on the recognition of speech, with one and two ears,'' *J. Acoust. Soc. Am.* 25,975–979.

[2] Bronkhorst, A. (2000). ''The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,'' *Acustica*. 86, 117–128.

[3] Egan, J., Carterette, E., and Thwing, E. (1954). ''Factors affecting multi channel listening,'' *J. Acoust. Soc. Am.* 26, 774–782.

[4] Dirks, D., and Bower, D. (1969). ''Masking effects of speech competing messages,'' *J. Speech Hear. Res.* 12, 229–245.

[5] Festen, J., and Plomp, R. (1990). ''Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing,'' *J. Acoust. Soc. Am.* 88, 1725–1736.

[6] Darwin, C., and Hukin, R. (2000). ''Effectiveness of spatial cues, prosody and talker characteristics in selective attention, *J. Acoust. Soc. Am.* 107, 970–977.

[7] Brungart, D. (2001a). ''Informational and energetic masking effects in the perception of two simultaneous talkers,'' *J. Acoust. Soc. Am.* 109, 1101– 1109.

[8] Brungart, D. (2001b). ''Informational and energetic masking effects in the perception of multiple simultaneous talkers,'' *J. Acoust. Soc. Am.* 110, 2527– 2538.

[9] Saberi, K. and Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, 398, 760.

[10] New, B., Pallier, C., Brysbaert, M., Ferrand, L. Lexique 2: A New French Lexical Database (In Press) *Behavior Research Methods, Instruments, & Computers.*

Index des auteurs

Index thématique

Phonétique et phonologie

Production de parole

Prosodie