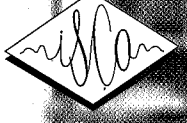
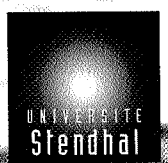
Actes des XXIIIèmes Journées d'Etude sur la Parole

JEP'2000

Aussois - France
19-23 juin 2000





XXIIIèmes Journées d'Etude sur la Parole

JEP'2000

Centre Paul Langevin
Aussois - Savoie

19 - 23 Juin 2000

Organisées par
l'Institut de Communication Parlée (ICP)

Pour
le Groupe Francophone de la Communication Parlée (GFCP)
de la Société Française d'Acoustique (SFA)
et de l'International Speech Communication Association (ISCA)

Avec le soutien
du CNRS, de l'Université Stendhal, de l'Institut National Polytechnique de Grenoble (INPG) et de la fédération ELESA,
de la ville de Grenoble et du Conseil de la Région Rhônes-Alpes,
de la Délégation Générale à la Langue Française (DGLF) et du Centre National d'Etudes en Télécommunications de France Télécoms (CNET)

1 Présentation

1.1 Objectifs

Les JEP'2000 sont consacrées aux travaux couvrant l'ensemble des aspects fondamentaux de la communication parlée dans ses diverses modalités ainsi que les différentes problématiques liées à son traitement automatique. Une attention particulière sera apportée aux recherches qui visent à la compréhension et/ou à la simulation des processus cognitifs en jeu lors de la production et/ou la perception de parole.

1.2 Comité scientifique

M. ADDA-DECKER, LIMSI, Paris (F) R. ANDRÉ-OBRECHT, IRIT, Toulouse (F) P. BADIN, ICP, Grenoble (F) G. BAILLY, ICP, Grenoble (F) F. BIMBOT, IRISA, Rennes (F) J.-F. BONASTRE, LIA, Avignon (F) J. CAELEN, CLIPS, Grenoble (F) M.-J. CARATY, LAFORIA, Paris (F) J.-L. COCHARD, SWISSCOM, Berne (CH) P. DELÉGLISE, LIUM, Le Mans (F) P. DUPONT, EURISE, St-Etienne (F) W. HESS, Univ. de Bonn (G) J.-M. HOMBERT, DDL, Lyon2 (F) Y. LAPRIE, CRIN, Nancy (F) P. PERRIER, ICP, Grenoble (F) J. SCHOENTGEN, ULB, Bruxelles (B) R. SOCK, IP, Strasbourg (F) B. TESTON, LPL, Aix-en-Provence (F) J. VAISSIÈRE, IPP, Paris (F)

1.3 Comité d'organisation

P. BADIN, G. BAILLY (Présidents)
C. PULFER (Secrétaire), C. BULFONE (Administration électronique)

1.4 Le mot du comité d'organisation

Les JEP'2000 constituent un espace d'échanges unique pour la communauté francophone travaillant sur la parole. Cet espace d'échanges pluri-displinaires réunit tous les deux ans plus d'une centaine de chercheurs dont plus de la moitié sont de jeunes doctorants. Traditionnellement, les JEP ne proposent pas de sessions en parallèle: ainsi tous les participants ont la possibilité d'effectuer un tour d'horizon complet sur les disciplines impliquées dans l'étude de la parole. Les conférences invitées permettent de plus d'offrir un cadre introductif adapté à ce large public. Les JEP ont donc l'aspect formel d'un congrès international - avec un processus de sélection permettant au comité scientifique d'effectuer un vrai travail d'évaluation mais aussi de conseil - avec la convivialité d'un séminaire. Afin de favoriser encore plus l'échange et les ponts inter-disciplinaires, nous avons opté cette année pour un format original rassemblant en un même lieu relativement isolé l'hébergement et les communications. Le centre Paul Langevin du CAES à Aussois offre cette synergie dans les meilleures conditions. Nous espérons que les participants sauront mettre à profit cette occasion unique pour resserrer les liens scientifiques et souvent amicaux qui unissent notre communauté.

Nous tenons à remercier tous nos partenaires institutionnels: le CNRS, l'Université Stendhal, l'Institut National Polytechnique de Grenoble et la fédération ELESA. Les collectivités locales, la ville de Grenoble et le Conseil de la Région Rhônes-Alpes, ainsi que nos partenaires, la Délégation Générale à la Langue Française (DGLF), le Centre National d'Etudes en Télécommunications de France Télécoms (CNET) nous ont également apporté un soutien financier. Grâce à ce montant exceptionnel de subventions et une participation du GFCP, un nombre important de participants, notamment nos collègues non français, ont pu ainsi bénéficier de bourses sans lesquelles ils n'auraient vraisemblablement pas pu participer à cette rencontre.

Bien sûr, tout cela n'aurait pas pu se réaliser sans le dévouement de Corinne Pulfer et des nombreuses personnes qui ont géré l'administration de ce congrès, en particulier Joëlle Miguet et tout le service comptable de la délégation "Alpes" du CNRS.

N'oublions pas le travail remarquable du comité scientifique et de tous les chercheurs qui les ont aidés à évaluer et à améliorer les versions initiales des papiers présentés dans cet ouvrage. Merci enfin et surtout aux auteurs et à nos conférenciers invités qui ont fourni la matière de ces quelques 450 pages.

Bon séjour dans nos Alpes savoyardes,
Gérard Bailly & Pierre Badin

1.5 Message du Président du GFCP

Deux ans après Martigny, c'est à Aussois que nous nous retrouvons pour cette XXIIIème édition des Journées d'Etude sur la Parole (ou JEP'2000), organisées comme à l'habitude, sous l'égide du Groupe Francophone de la Communication Parlée (GFCP) de la Société Française d'Acoustique (SFA), également Groupe Spécialisé de l'ISCA (International Speech Communication Association).

Comme vous le savez, c'est à l'Institut de la Communication Parlée (ICP) qu'incombe l'organisation de ces JEP'2000, et je peux témoigner de la très vive motivation de ce laboratoire, exprimée de longue date, pour être l'hôte de ces premières JEP du millénaire. Cette motivation s'est également traduite au fil des mois, par les efforts déployés par le Comité d'Organisation pour mettre en place tous les éléments indispensables à la réussite de cette conférence, pour laquelle nous nous retrouvons aujourd'hui.

Année après année, les JEP ont conservé leur vocation d'être un point de rencontre sans équivalent. Point de rencontre entre les multiples disciplines de la communication parlée. Point de rencontre également entre les pays et les communautés de la francophonie. Cette double diversité est une richesse que nous avons tous envie de préserver, car elle nous permet de maintenir une vision d'ensemble de notre domaine tout en échangeant nos idées dans la langue dans laquelle nous sommes le plus à l'aise pour nous exprimer et pour nous comprendre.

Une autre spécificité des JEP est la forte proportion de communications présentées par des doctorants, qui font souvent leur premières armes à cette occasion. En ce sens, c'est une étape intermédiaire entre les présentations décontractées des Rencontres Jeunes Chercheurs (RJC) et la participation à des conférences internationales de grande taille et plus impressionnantes, comme Eurospeech ou ICSLP.

La procédure de sélection des communications aux JEP, mise en place depuis plusieurs années, est maintenant bien rodée et semble donner entière satisfaction. Rappelons qu'elle est basée sur une sélection à partir d'un article complet avec possibilité de modification de la première version de l'article, sur demande des relecteurs. Cette façon de procéder nécessite un effort supplémentaire de la part des auteurs (et des relecteurs) mais elle permet une meilleure interactivité, donnant ainsi un caractère plus constructif à l'ensemble du processus de sélection. J'en profite d'ailleurs pour remercier, au nom du GFCP, tout ceux qui ont accepté d'y participer.

Une centaine de communications figurent au programme de ces JEP'2000, couvrant des thèmes aussi divers que la phonologie, la synthèse de parole, la production, les modèles de langage, etc... Notons quelques thèmes émergents comme "indexation et analyse de scènes sonores" ou bien "parole et cognition" qui donnent lieu à deux conférences invitées et deux sessions orales en ouverture et en fermeture de ces JEP. Les autres conférences invitées traitent de la prosodie d'une part et de la reconnaissance de parole, d'autre part. L'ensemble du programme rend donc parfaitement compte de la pluridisciplinarité de notre domaine.

Signalons aussi la présence d'un créneau réservé à une session spéciale sur la question des expertises d'identification vocale à des fins judiciaires. Cette problématique occupe notre communauté scientifique depuis plusieurs années. et, encore récemment, la SFA et le GFCP ont dû réagir à des cas concrets d'utilisation de ces techniques dans le cadre d'affaires judiciaires. Un mois avant ces JEP, le GFCP a lancé une vaste consultation électronique, auprès (et au-delà) de la communauté scientifique en acoustique pour dégager les points à aborder en priorité et identifier différentes solutions susceptibles de faire évoluer la situation. Ces éléments seront débattus pendant la session spéciale, à laquelle nous espérons que vous serez nombreux à participer, car il s'agit là d'une problématique touchant directement à une question de société sur laquelle notre implication est indispensable.

Dans les mois prochains se tiendront les élections pour le renouvellement du Comité du GFCP, à l'occasion desquelles vous serez amenés à élire 15 membres de notre groupe spécialisé. Plusieurs membres du Comité actuel quitteront leur fonction et il faut dès maintenant préparer la relève. Nous espèrons que vous serez nombreux à vous porter candidats. Un point relatif à cette question figure à l'ordre du jour de l'Assemblée Générale du GFCP qui aura lieu pendant la conférence. N'hésitez pas à nous contacter pour de plus amples informations.

Le GFCP est également un Groupe Spécialisé de l'ISCA (anciennement ESCA). Cet ancrage du GFCP à une association internationale en communication parlée complète bien notre affiliation historique à la communauté scientifique francophone en acoustique. Cela nous permet tout à la fois de promouvoir notre domaine auprès de nos collègues francophones spécialisés dans des disciplines voisines de la nôtre et d'exprimer la sensibilité francophone auprès de nos homologues internationaux travaillant directement dans le domaine de la communication parlée.

Un de nos rôles au sein de l'ISCA doit consister à promouvoir la préservation de la diversité, et démon-

trer qu'elle peut et doit aller de pair avec le processus d'internationalisation. Je prendrai comme exemple la possibilité qui sera offerte en 2001, à titre expérimental, d'adjoindre aux articles pour Eurospeech un second résumé dans une autre langue que l'anglais, pour inclusion sur les actes au format CD-ROM. Ceci permettra notamment un pointage plus efficace des moteurs de recherche multilingue sur les articles publiés et un accès plus immédiat au contenu de ces publications pour nos collègues francophones. Soyons donc nombreux à utiliser cette possibilité qui nous est offerte, pour indiquer que nous souhaitons que cette disposition soit maintenue et généralisée à d'autres conférences.

Mais revenons aux JEP'2000 qui s'ouvrent maintenant. Et avant de laisser la place aux échanges scientifiques, je voudrais adresser une nouvelle fois mes remerciements, ou plutôt nos remerciements, au Comité d'Organisation, et plus particulèrement à Pierre BADIN, à Gérard BAILLY et à Corinne PULFER, qui n'ont pas ménagé leurs efforts pour faire de ces JEP'2000 un millésime exceptionnel.

Je vous souhaite à tous une excellente conférence.

Frédéric BIMBOT Président du GFCP

Table des matières

Conférences invitées

Indexation, Segmentation et Analyse de Scènes

Phonétique/Phonologie

Synthèse

Production

Prosodie

Reconnaissance de la parole

Parole et cognition

Prosodie/phonologie

Reconnaissance et modèles de langage

Reconnaissance de la langue et du locuteur

Production et pathologies

Liste alphabétique des auteurs

Conférences invitées

Analyse de scène auditive et parole

Alain de Cheveigné

Ircam - CNRS
1 place Igor Stravinsky, 75004, France
Tél.: ++33 1 44 78 48 46 - Fax: ++33 1 44 78 15 40
Mél: cheveign@ircam.fr - http://www.ircam.fr/pcm/cheveign

ABSTRACT

Auditory Scene Analysis (ASA) is a more basic competence than speech perception, phylogenetically more ancient, but the two share important relations. Speech communication often occurs in presence of interfering noises and voices, and therefore depends on segregation mechanisms for reception. Speech has often been used as a stimulus to investigate ASA phenomena, but in some respects it appears to escape from basic principles. The most interesting potential application of computational auditory scene analysis (CASA) is in speech recognition systems.

1. INTRODUCTION

Pour guider son action et faciliter sa survie, l'organisme construit un modèle du monde qui l'entoure à l'aide d'informations fournies par ses sens. Lorsque l'environnement est complexe, cette information doit être triée et distribuée parmi les éléments du modèle. Si le tri se fait bien, le modèle est fidèle, l'action efficace, et les chances de survie bonnes. Dans le domaine de l'audition, cette opération s'appelle *analyse de scène auditive* (ASA). L'analyse de scène a été pratiquée par nos ancêtres avant qu'ils se mettent à parler, et on peut penser que ses mécanismes sont pour l'essentiel génériques et non spécifiques à la parole. ASA et parole entretiennent néanmoins des rapports privilégiés pour plusieurs raisons, qui tiennent à l'importance que revêt pour nous la communication parlée.

• Même si la parole n'était pas le stimulus le plus important pour nos lointains ancêtres, elle l'est devenue pour nous. La voix d'un locuteur est souvent l'objet des processus ASA, et celle d'un locuteur concurrent une source de bruit masqueur.

• La parole semble paradoxalement échapper à certains principes qui régissent l'ASA de sons plus simples. Par exemple on s'attendrait à ce que l'irrégularité des sons qui composent la parole s'oppose à sa fusion en un flux unique, mais il n'en est rien.

• La parole (plus ou moins stylisée) a souvent été utilisée comme matériau d'expérience pour explorer l'ASA. L'identification des voyelles concurrentes constitue notamment un paradigme puissant pour l'étude des processus d'organisation simultanée.

• Des applications importantes relèvent de la parole: reconnaissance de la parole en milieu bruité, prothèses auditives, implants cochléaires. C'est ainsi que s'est développée la discipline d'*analyse de scène auditive computationnelle* (CASA), qui cherche à reproduire les opérations de l'ASA par des moyens computationnels.

L'article est en trois parties. La première résume brièvement quelques principes de l'ASA. La deuxième discute du rôle du voisement, un indice de ségrégation parmi les plus importants. La troisième passe en revue quelques tentatives d'application de l'analyse de scène computationelle à la reconnaissance de la parole.

2. LES PRINCIPES DE L'ASA

Jusqu'à une époque récente, l'Audition s'intéressait à la perception de qualités telles que la hauteur, la sonie, le timbre, etc., d'un son émis par une *source unique*. La phonétique décrit de même les propriétés acoustiques d'une voix isolée. En pratique nous percevons souvent les sons parmi une cacophonie de voix et bruits superposés. Chaque oreille reçoit des ondes provenant d'une multitude de sources, mais on peut souvent porter son attention sur une source particulière et juger de sa sonie, de sa hauteur, de son timbre, voire comprendre ce qui est dit lorsqu'il s'agit de parole. Les modèles classiques, conçus pour traiter une source isolée, ne sont pas suffisants pour expliquer la perception dans ce cas.

Helmholtz [Hel87] déjà se demandait comment on pouvait percevoir les qualités individuelles des instruments de l'orchestre, mais il a fallu attendre le travail de Bregman [Bre90] pour que l'analyse de scène auditive devienne un sujet d'étude à part entière. Pour Bregman, le problème de l'émergence de sources subjectives (flux, ou "streams") est principal, et la détermination de leurs qualités secondaire, puisque logiquement le premier est un préalable au second. Pour élaborer sa théorie, Bregman s'est appuyé sur l'analogie avec l'analyse de scène en vision, et les principes de la psychologie Gestalt.

Figure 1. Analyse de scène visuelle. A gauche, les fragments paraissent inorganisés. A droite, la présence d'une forme masquante permet leur regroupement perceptif. L'ASA cherche des principes analogues pour l'organisation du monde sonore. (D'après [Bre90]).

Avec le développement de l'Informatique et de l'Intelligence Artificielle sont apparues des tentatives d'*Analyse de Scène Auditive Computationnelle* (CASA) [Lyo83, Wei85, Coo91, Mel91, Bro92, Wan92, Ell96]. Les modèles CASA ont la double ambition d'aider à comprendre les processus perceptifs, et de résoudre des problèmes pratiques, par exemple éliminer le bruit dans un système de reconnaissance de la parole. L'influence de la vision computationnelle, notamment les travaux de Marr [Mar82], a joué un rôle déterminant dans l'élaboration de ces modèles.

Les mots de l'ASA

Il faut distinguer la *source acoustique* de l'entité perceptive qui lui correspond après analyse, qu'on désigne par *flux* ("stream") ou par les termes plus ambigus d'*objet* ou *événement* perceptif. On oppose la *fusion* (groupement) à la *scission* (ségrégation) selon que l'information acoustique évoque un ou plusieurs flux. On parle d'organisation *simultanée* ou *séquentielle* selon que les sources à analyser se manifestent de façon simultanée ou séquentielle. Enfin, on distingue les processus *primitifs* (montants, ou "bottom-up") des processus *à base de schémas* (descendants, ou "top-down").

Organisation simultanée

Il arrive que des sources distinctes se manifestent en même temps (ou avec un chevauchement temporel). Les corrélats acoustiques qui parviennent aux oreilles sont intimement mêlés, mais malgré cela nous entendons parfois plusieurs objets perceptifs, correspondant chacun à une source, plutôt que le flux unique qu'évoquerait le son d'une source unique. Quels aspects du signal acoustique font qu'il évoque la perception d'un objet plutôt que deux ?

Un premier facteur de cohésion est la *simultanéité d'attaque*, et plus généralement la communauté de variation des composantes fréquentielles du son. Lorsque toutes démarrent en même temps on tend à percevoir une source unique. Une asynchronie d'attaque évoque la perception d'objets multiples. C'est un exemple du principe plus général de *destin commun*.

Un deuxième facteur important est l'*harmonicité*. Lorsque les partiels suivent une série harmonique unique (sons de certains instruments, parole voisée) le son évoque une source unique. Dans le cas contraire (la "polypériodicité" de Marin [Mar91]), le stimulus paraît provenir de plusieurs sources.

Un troisième facteur est la *corrélation binaurale*. Si les composantes du son ont toutes la même relation binaurale, leur fusion est favorisée. Des disparités entre composantes peuvent évoquer des sources multiples, distinctes dans l'espace.

L'organisation auditive va au-delà d'une simple perception de multiplicité, puisque dans une certaine mesure une source parmi des sources concurrentes est perceptible comme si elle était isolée. Pour des sources qui se chevauchent dans le temps, la ségrégation perceptive est gouvernée par la règle *ancien plus nouveau* (old plus new). A l'apparition d'une nouvelle source, le spectre est analysé en défalquant la contribution estimée de l'ancienne (supposée toujours présente). Bien entendu, un tel mechanisme doit savoir faire la différence entre les variations de spectre dues à l'apparition d'une nouvelle source, et celles qui tiennent de la nature intrinsèquement variable d'une source unique.

La structure harmonique de sources concurrentes est aussi exploitée pour la ségrégation, comme on le verra en détail plus loin. Bregman [Bre90] donne d'autres exemples de traits qui gouvernent l'analyse simultanée. La *modulation de fréquence* a souvent été évoquée comme un exemple du principe de destin commun. Pour comprendre cette notion, il suffit d'imaginer une représentation spectro-temporelle de façon graphique. Des composantes dont la modulation est cohérente devraient former une "figure", et se distinguer de composantes immobiles ou dont la modulation serait incohérente. L'idée est attrayante, mais on verra qu'en fait cet indice joue un rôle mineur.

Organisation séquentielle

Il arrive que des sources se manifestent de façon répétée dans le temps, et évoquent la perception d'une suite d'entités cohérente (flux), distincte des autres sons de l'environnement. C'est le cas d'une voix, d'une ligne mélodique, d'une succession de crissements de pas dans la neige, etc. Quels aspects d'une succession de sons font qu'ils se groupent en un flux unique, plutôt qu'en plusieurs flux parallèles ?

Un premier facteur est la similitude. Des sons disparates tendent à former des flux multiples, alors que des sons proches forment un flux unique. Il peut s'agir d'une similitude de hauteur, de timbre, de position spatiale, de sonie, etc., mais le facteur le plus important est la similitude des activités évoquée dans le système auditif périphérique [Har91] et donc du contenu spectral.

Un deuxième facteur est la vitesse de présentation des sons. Un taux lent favorise la formation d'un flux unique,

alors qu'un taux rapide favorise la scission en des flux multiples.

Bregman interprète ces facteurs en termes de principes Gestalt tels que celui de destin commun (ou origine commune): les sons d'une même source ont toutes les chances d'être semblables, ou de varier lentement, et il est naturel d'interpréter une variation importante et/ou rapide comme l'intervention d'une nouvelle source. Bregman note cependant le paradoxe de la parole, dont les variations (notamment les transitions consonantiques) sont grandes et rapides, sans que cela compromette la cohérence de la voix.

Fusion vs scission

La psychoacoustique classique considère des sources uniques et leur attribue l'ensemble des représentations de bas niveau (physiologiques) et de haut niveau (perceptives) qu'elles évoquent. L'ASA suppose des sources multiples et donc une *décomposition* des représentations. On peut imaginer que cette décomposition a lieu selon des dimensions tonotopiques, "périodotopiques", "spatiotopiques", etc., dans des cartes supposées exister dans le système auditif central. Si une telle décomposition (scission) est possible, alors il faut expliquer pourquoi elle ne se fait pas systématiquement, c'est-à-dire qu'il faut expliquer la cohésion dans le cas d'une source unique. Fusion et scission sont les deux faces d'une même pièce.

Processus primitifs vs schémas

Bregman distingue les processus d'organisation primitifs, qui font intervenir des mécanismes ascendants ("bottom-up"), de ceux qui font intervenir des attentes produites par le contexte, ou des "schémas" présents chez l'auditeur, et qui pourraient faire intervenir des processus descendants ("top-down"). Les processus primitifs joueraient pour tout type de son, alors que les schémas seraient spécifiques de sons particuliers, par exemple la parole. Les schémas qui interviennent dans le décodage de la parole pourraient ainsi jouer un rôle dans l'organisation auditive de scènes comprenant une voix. Cependant il est difficile de démontrer que le schéma est intervenu dans le processus d'organisation lui-même, plutôt que dans une phase ultérieure d'interprétation des données organisées.

Un exemple souvent cité est celui de la "restauration phonémique" [War70]. Un segment phonémique de parole est excisé et remplacé par un bruit, mais l'auditeur croit entendre le segment manquant. Plus étrange, il est incapable de situer avec précision l'interruption au sein du mot. Lorsque l'interruption introduit une ambiguïté, la restauration peut dépendre du contexte qui la précède (ou même qui la suit). On voit parfois dans ce phénomène l'indice d'un mécanisme de bas niveau de partition de l'information acoustique, ou de synthèse de la partie manquante, à partir d'un schéma. On peut aussi l'attribuer plus prosaïquement au mécanisme d'*interprétation* des données présentes, exploitant le fait que la présence de la parole derrière le bruit est une hypothèse plus probable que celle de son absence.

Outre les processus à base de schémas, la parole devrait aussi bénéficier des processus primitifs génériques applicables à tout son, mais l'hypothèse a aussi été défendue que la parole est "spéciale" et leur échappe [Rem94].

3. VOISEMENT ET SÉGRÉGATION

Les aspects de l'ASA qui relèvent de la parole (et inversement) sont nombreux. Plutôt que d'entreprendre une revue générale qu'on peut trouver ailleurs [Bre90, CoE00], je développerai un aspect de l'ASA que j'ai exploré plus en profondeur.

Une particularité de la voix humaine est qu'elle comprend des portions *voisées* à la structure approximativement périodique ou harmonique. Le voisement joue un rôle bien connu comme vecteur d'informations prosodiques liées à la fréquence de vibration des cordes vocales (F0). Cherry [Che53] a suggéré qu'une différence de F0 moyenne entre des voix d'homme et de femme pourrait aussi faciliter la compréhension lorsque ces voix sont superposées. La F0 participerait au fameux "effet cocktail" (cocktail party effect). Brokx et Nooteboom [Bro82] ont confirmé cette intuition, en montrant que l'intelligibilité est meilleure lorsque les F0 sont dans des plages différentes plutôt que superposées. On peut interpréter ce résultat de plusieurs façons.

Une première interprétation est que la F0 sert à "suivre" une voix au cours du temps. La F0 décrit une courbe relativement continue dans les parties voisées, et entre parties voisées les variations sont limitées et partiellement prévisibles. Dans une expérience de Darwin [Dar75] des auditeurs devaient répéter les mots d'une voix présentée à une oreille en ignorant une voix parasite dans l'autre. Lorsque la F0 fut brusquement intervertie entre les voix, les sujets répétèrent quelques mots de l'oreille opposée, en suivant donc la continuité de la F0. Dans une autre expérience, un changement brusque de F0 dans une transition entre voyelles évoquait la perception de voix multiples prononçant des consonnes [Dar77].

Une deuxième interprétation des résultats de Brokx et Nooteboom est qu'une différence de F0 (ΔF0) facilite la ségrégation dans les parties voisées qui se chevauchent. Pour tester cette interprétation, Scheffers [Sch83] a assemblé des mélanges de voyelles synthétiques concurrentes, et demandé à des sujets de les identifier. L'identification était meilleure lorsque les F0 des voyelles étaient différentes, ce qui confirme donc aussi cette deuxième interprétation. Par la suite, de nombreux auteurs ont utilisé le paradigme des "voyelles doubles" pour tenter de comprendre les mécanismes de ségrégation (Figure 2).

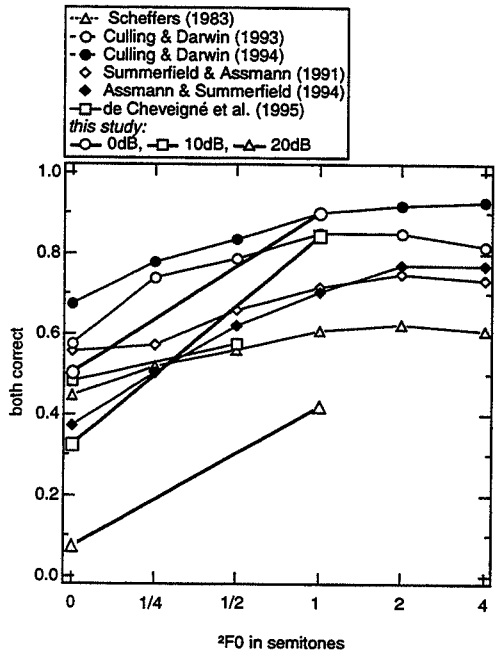


Figure 2. Taux d'identification de paires de voyelles en fonction de la différence entre leurs F0. L'identification profite de mécanismes de ségrégation qui exploitent la structure harmonique des voyelles [deC97a].

Une hypothèse est que la *structure harmonique* de la parole voisée est exploitée par un mécanisme de ségrégation. Les partiels des deux voix suivent des séries harmoniques distinctes, ce qui pourrait faciliter leur tri (sauf bien sûr à $\Delta F0=0$). Par exemple Parsons [Par76] a proposé une méthode de séparation de voix fondée sur des spectres de Fourier calculés sur des fenêtres de 51.2 ms. Les pics du spectre étaient regroupés dans des séries harmoniques, et attribués à l'une ou l'autre voix. Scheffers [Sch83], puis Assmann et Summerfield [AsS90] s'en sont inspirés pour élaborer un modèle de ségrégation fondé sur l'analyse fréquentielle de la cochlée, mais ils ont constaté une résolution trop faible. La sélectivité *fréquentielle* cochléaire étant insuffisante pour isoler les partiels de chaque série, les modèles plus récents s'orientent vers des mécanismes *temporels* (voir plus loin).

A regarder de plus près, pour extraire une voix on dispose de *deux* structures harmoniques: celle de la voix cible et celle de sa concurrente. Servent-elles toutes les deux, ou seulement une, et si oui, laquelle ? Deux mécanismes sont possibles: le *renforcement harmonique* par lequel une voix est favorisée par son harmonicité, ou l'*annulation harmonique*, par laquelle elle est favorisée par l'harmonicité de son concurrent. Le premier est attrayant car il s'applique quel que soit le bruit, périodique ou non, mais il n'est utile que pour les parties voisées de la cible. L'annulation harmonique, elle, marche pour les parties voisées et non voisées, mais seulement si le concurrent est harmonique. Zissmann et Weinstein [ZiW83] ont simulé ces deux stratégies avec de la parole mélangée, en supprimant la voix concurrente soit lorsque cette voix était voisée, soit lorsque la cible était voisée. L'intelligibilité était meilleure dans le premier cas, ce qui indique que la stratégie d'annulation est plus utile (ou le serait si on pouvait implémenter les deux stratégies de façon parfaite). Du fait des apériodicités de la parole naturelle l'implémentation de l'une et l'autre stratégie est forcément imparfaite, mais on peut montrer que cela affecte moins l'annulation que le renforcement [deC93].

On peut utiliser le paradigme expérimental de Scheffers pour connaître la stratégie utilisée par le système auditif, à condition de mesurer séparément l'identification de chaque voyelle d'une paire (le paradigme classique compte les réponses pour lesquelles les voyelles sont simultanément correctes). Pour des mélanges de voyelles voisées et chuchotées, Lea [Lea92] a constaté que la ségrégation bénéficiait à la seule composante chuchotée. D'autres auteurs ont confirmé ce résultat en montrant que le facteur qui détermine la ségrégation est l'*harmonicité de la voyelle concurrente* [SuC92a, deC95, deC97a, deC97b]. L'harmonicité propre de la cible ne lui est d'aucun secours, résultat d'autant plus étonnant que plusieurs algorithmes de séparation de parole et modèles CASA utilisent l'harmonicité de la voix cible.

Les modèles et méthodes peuvent se classer selon qu'ils adoptent la stratégie de renforcement ou celle d'annulation (que favorisent les arguments précédents) [deC93a, deC95]. Parmi les modèles d'annulation, celui de Meddis et Hewitt [MeH92] est le mieux connu. Ce modèle fait un tri parmi les canaux périphériques issus de la cochlée. Les trains d'impulsions du nerf auditif sont soumis à un processus d'autocoincidence neuronale selon le modèle de Licklider [Lic56] qui permet de mesurer leur périodicité. Les canaux dont la périodicité ne correspond *pas* à celle qui domine le stimulus sont regroupées et utilisées pour extraire la voix la plus faible. Alors que le modèle de Parsons/Scheffers nécessitait une résolution spectrale à l'échelle des *partiels*, celui de Meddis et Hewitt nécessite seulement une résolution à l'échelle des *formants*. Une différence des structures formantiques fait que différents canaux sont dominés par différentes voix, même lorsque les voix sont de même amplitude.

En revanche, si les amplitudes sont suffisamment différentes, la voix la plus forte risque de dominer *tous* les canaux périphériques, dans quel cas ce mécanisme de ségrégation ne marche pas et on ne prévoit pas d'effet de $\Delta F0$. Cela fournit un moyen de tester le modèle de Meddis et Hewitt. Les premières expériences "voyelles doubles" utilisaient des voyelles égalisées en amplitude, sonie ou "force d'excitation" (selon l'expérience), mais des expériences plus récentes ont introduit des différences de niveau entre voyelles [deC99b]. L'effet de $\Delta F0$ est plutôt renforcé pour la composante faible, et reste important jusqu'à -25 dB (Figure 3). Dans ces conditions, si on modélise le pattern de dominance des canaux périphériques, on s'aperçoit que *tous* sont dominés par la voyelle la plus forte. Le modèle de Meddis et Hewitt ne peux donc pas expliquer ces effets de $\Delta F0$.

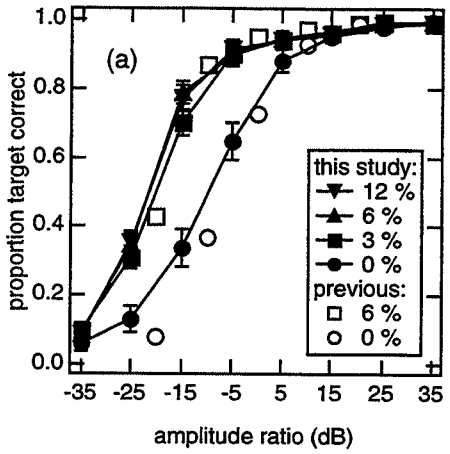


Figure 3. Taux d'identification en fonction de l'amplitude d'une voyelle relative à la sa concurrente, pour des F0 égaux (cercles) ou différents (autre symboles) [deC99b]

A côté de modèles spectraux [Sch83, Par76] et spectro-temporels [MeH92, AsS90], on peut imaginer des modèles purement *temporels* de ségrégation harmonique. Le *filtre d'annulation neuronal* est un circuit hypothétique comprenant un neurone "porte" muni de deux synapses, l'un excitateur alimenté directement, et l'autre inhibiteur alimenté via une ligne à retard. Toute impulsion arrivant par le chemin direct est transmise, sauf si une impulsion arrive simultanément par le chemin retardé. L'effet du filtre est de modifier la statistique des intervalles inter-impulsions, et il se trouve que cela suffit à supprimer la contribution d'une source dont la période serait égale au retard [deC93]. A l'aide de ce filtre on peut constituer un modèle de perception de voyelles concurrentes qui explique l'ensemble des données expérimentales [deC97, deC99d]. Il peut aussi servir de base à des modèles de perception de la hauteur [deC98, deC99a, deC99f].

Avant qu'on ne découvre le peu d'effet de l'harmonicité d'une cible, on pouvait imaginer que la structure harmonique pourrait grouper ensemble les formants d'une voix, en les étiquetant d'une même périodicité. Culling et Darwin [CuD93] ont présenté à des sujets des paires de voyelles synthétisées de telle façon que le formant F1 de l'une ait la même périodicité que les formants supérieurs de l'autre. Cette manipulation eut peu d'effets sur la ségrégation, ce qui suggère le manque d'efficacité d'une harmonicité commune pour grouper ensemble les formants d'une voyelle. Ces résultats et d'autres [CuD94] suggéraient même que l'harmonicité pourrait ne pas avoir du rôle du tout, puisqu'ils révélaient une ségrégation entre des voyelles dont ni l'une ni l'autre n'est parfaitement harmonique. Cela amena ces auteurs à proposer que, pour des ΔF0 petits (<6%) la ségrégation pourrait être due à un simple effet de battements [CuD94, AsS94], sans aucun lien avec la structure harmonique. Cette hypothèse a l'attrait de ne pas nécessiter une résolution fréquentielle fine. Une hypothèse apparentée, proposée à la même époque (l'hypothèse PPA de [AsS94]), est que la structure temporelle particulière de la période de voisement (impulsion glottique suivie d'un intervalle d'énergie faible) fournirait une "fenêtre" qui faciliterait la perception de la deuxième voyelle.

Ces deux hypothèses (battements et PPA) impliquent, si elles sont vraies, une certaine dépendance sur la *phase* des voyelles. Par exemple une manipulation de phase qui chamboule la structure temporelle intra-période devrait mettre en échec le mécanisme PPA. Deux séries d'expériences [deC97b, deC99d] ont montré l'absence quasi-total d'effets du spectre de phase, ce qui permet d'écarter ces deux hypothèses, même pour des ΔF0 faibles. C'est seulement à des F0 très bas (50 Hz) qu'on constate des effets de l'alignement temporel entre périodes [AsS90].

Dans les expériences de voyelles doubles, l'identification croit en fonction de ΔF0 pour atteindre rapidement un plateau à partir d'environ 1/2 ton (6%). Elle décroît ensuite à l'octave, un phénomène qu'avait déjà noté Brokx et Nooteboom. En dessous de 6% l'effet de ΔF0 décroît, mais reste mesurable jusqu'à 0.4%, ou 1/16e de ton [deC99d]! Ce résultat indique la très grande finesse de résolution fréquentielle ou temporelle du mécanisme de ségrégation. Par exemple si on retient le modèle d'annulation neuronal, il doit pouvoir exploiter des disparités de l'ordre de 30 µs.

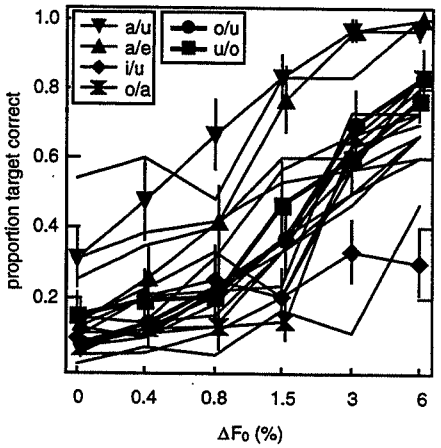


Figure 4. Taux d'identification en fonction de ΔF0 pour différentes paires de voyelles (20 combinaisons des 5 voyelles du japonais). Une ΔF0 de 0.4% suffit à produire un effet significatif [deC99d].

Des ΔF0 non-nuls facilitent la ségrégation, mais de nombreux auteurs ont noté avec étonnement que l'identification à ΔF0=0 dépasse de loin le hasard. Le spectre du stimulus est pourtant très différent de l'une ou l'autre voyelle. Pour mieux comprendre les facteurs qui déterminent l'identification, j'ai utilisé des paires de voyelles synthétisées avec le même spectre de phase (pour simplifier la sommation), et une large plage d'amplitudes relatives. L'analyse des résultats par paire a montré qu'une voyelle est identifiée dès lors que ses formants F1 et F2 sont saillants (les formants supérieurs semblent moins importants sans qu'on puisse écarter tout rôle). L'identification n'est pas perturbée par la saillance des formants du concurrent. Les deux voyelles peuvent d'ailleurs partager un formant: le principe d'allocation exclusive [Bre90] ne joue donc

pas. Ces conclusions furent tirées à ΔF0=0. L'effet d'une ΔF0 non-nul est apparemment de renforcer encore la saillance des indices formantiques.

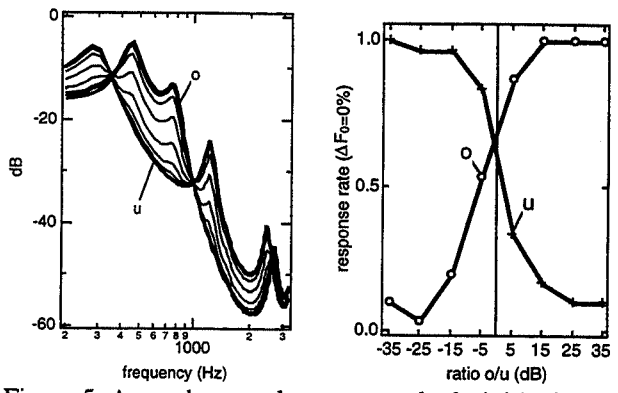


Figure 5. A gauche: enveloppe spectrale de /o/ (trait pointillé gros), de /u/ (trait continu gros) et de stimulis hybrides /o/+/u/ pour des rapports d'amplitudes allant de +35 à -35 dB par pas de 10 dB (traits fins). A droite: taux d'identification des voyelles /o/ et /u/ pour ces stimulis hybrides à ΔF0=0. L'analyse des données pour l'ensemble des paires de voyelles permet de comprendre quels indices supportent l'identification des voyelles [deC99a].

La *largeur de bande des formants* est connue pour avoir peu d'effet sur le timbre ou l'identité d'une voyelle. En revanche elle se révèle jouer un rôle important dans la ségrégation. Une voyelle à formants étroits est plus masquante, et plus résistante au masquage, qu'une voyelle à formants larges [deC99e], et cet effet s'ajoute aux effets éventuels de la ΔF0. Dans la mesure où les formants s'affinent lorsqu'on raccourcit la phase d'ouverture de la glotte, ce facteur pourrait varier sous contrôle musculaire et jouer ainsi un rôle actif dans les situations de compétition verbale (cette hypothèse reste à vérifier).

Un facteur qui s'apparente à la ΔF0 est la *modulation de fréquence* (FM), souvent citée comme exemple du principe Gestalt de destin commun [Bre90]. Une modulation commune des partiels d'un son permettrait (selon ce principe) leur regroupement, et leur ségrégation d'avec les partiels d'un fond statique ou modulé de façon différente. Des essais informels ont montré des effets spectaculaires, parfois repris dans des démos, mais dans chaque exemple la FM avait pour effet secondaire de créer une ΔF0 instantanée, suffisante pour expliquer à elle seule l'effet. Plusieurs études ont cherché à mettre en évidence un effet spécifique de la FM au delà des ΔF0 induits, sans succès [McA89, Dem90, SuC92b Car94, deC96, MaM97].

Pour résumer, l'harmonicité liée au voisement est l'un des facteurs les plus importants de l'analyse auditive de scènes comprenant de la parole. Les effets sont mesurables sur une large plage d'amplitudes (pour des voyelles jusqu'à -25 dB par rapport à leur concurrent), ce qui argue en faveur de leur utilité écologique. Des différences de F0 de 6% ou plus sont pleinement exploitables, et des effets sont mesurables pour des ΔF0 bien plus faibles. Cela confirme les intuitions des premiers chercheurs tels que Cherry. En revanche certains aspects sont étonnants et vont à l'encontre des intuitions et principes Gestalt. La structure harmonique de la voix cible ne joue guère de rôle dans sa ségrégation (sauf peut-être pour maintenir la continuité de la voix), et la modulation de fréquence a un effet faible ou nul.

4. CASA ET RECONNAISSANCE DE LA PAROLE

Notre système auditif exploite l'harmonicité liée au voisement pour améliorer l'identification d'une voix masquée par une voix concurrente. C'est l'un des mécanismes qui font que la perception d'un auditeur humain est plus robuste face au bruit que les systèmes de reconnaissance de la parole [Lip97]. Il est évident qu'il serait utile de reproduire ces mécanismes artificiellement. L'*analyse de scène auditive computationelle* (CASA) est une approche possible (pas la seule) pour y parvenir.

Weintraub [Wei85] le premier a tenté d'utiliser un modèle du système auditif pour améliorer la reconnaissance de la parole en présence d'une voix concurrente. Le modèle, proche de celui de Meddis et Hewitt [MeH92], s'appuyait sur une analyse par autocorrélation des canaux d'un banc de filtres. Weintraub travaillait à partir des idées de Lyon [Lyo83, Lyo84], qui lui-même s'inspirait de Licklider [Lic56]. Les canaux étaient assignés à une voix ou l'autre selon leur périodicité, puis la parole resynthétisée et présentée à un système de reconnaissance. Les taux de reconnaissance obtenus n'étaient pas excellents, mais le système de Weintraub a inspiré de nombreux efforts depuis.

Cooke [Coo91], Mellinger [Mel91], Brown [Bro92] et Ellis [Ell96] ont tous essayé de trouver des modèles physiologiquement plausibles capables d'exploiter le voisement, et d'autres indices, pour la ségrégation. Dans la plupart des cas, l'objectif fixé était la *resynthèse* de voix séparées. Cette objectif, qui correspond à l'idée naïve de "séparation" des voix, a l'avantage de permettre une évaluation immédiate par écoute ou mesure du rapport signal-sur-bruit [Bro92], mais on peut se demander s'il s'agit d'un objectif raisonnable. Une première remarque est qu'une resynthèse parfaite est impossible dans le cas général (du fait de l'indétermination du mélange). Un critère de qualité perceptive de la resynthèse risque donc d'être irréaliste. Une deuxième remarque est que la resynthèse n'a pas sa place dans un modèle du système auditif qui, lui, ne resynthétise pas. Une troisième remarque est que la resynthèse n'est ni nécessaire, ni forcément souhaitable pour une application de reconnaissance de la parole [Sla95, C0G00]. Certes, la resynthèse permet une architecture modulaire, mais une intégration plus étroite entre ségrégation et reconnaissance est souhaitable pour pleinement profiter de l'analyse de scène.

Cooke et ses collègues ont investi beaucoup d'effort dans cette question, en particulier en développant la *théorie des données manquantes* (TDM) pour gérer les données incomplètes fournies par un système CASA. Une ségrégation parfaite étant impossible, les données sont incomplètes, mais si les parties manquantes sont connues il est pos-

sible d'en tenir compte dans l'étape de reconnaissance [AhT93, CMG97, CGJ00, LiC97, deV99]. Deux méthodes sont proposées. L'une attribue un poids nul aux parties manquantes (méthode des "marginales"), l'autre les interpole à partir des données présentes à l'aide d'un modèle (méthode des "imputations"). La première est plus efficace, mais la deuxième a l'avantage de fournir des données "completes" (utiles par exemple pour une resynthèse éventuelle).

La TDM est sans doute une clé pour l'utilisation effective des modèles CASA pour la reconnaissance. Elle est d'une utilité plus large pour exploiter des données distordues ou de fiabilité inhomogène, dès lors que la distortion ou la fiabilité sont connues. C'est le cas par exemple de la distortion convolutive d'un réseau de microphones ou d'une analyse en composantes indépendantes. C'est le cas aussi pour la fusion de données multimodales [RoD98]. La TDM est proche par son esprit du modèle FLMP de Massaro [Mas90], et c'est un paradigme intéressant pour l'élaboration de modèles perceptifs [deC99c].

A noter que l'objectif de *reconnaissance* de la parole à la sortie d'un système CASA est plus réaliste que celui de *resynthèse* des voix séparées. Celle-ci est difficile du fait de l'indétermination notée plus haut, et aussi du fait de l'exigence des auditeurs. Une resynthèse de qualité n'est possible qu'à l'aide de modèles normatifs ou de production sophistiqués (dont les paramètres seraient contraints par les données incomplètes).

Pour résumer, la reconnaissance de la parole est l'application potentielle la plus intéressante des modèles d'analyse de scène. Il est certain que les systèmes profiteraient d'une robustesse semblable aux auditeurs humains. Inversement, les techniques de reconnaissance constituent l'un de nos meilleurs modèles pour la perception de la parole [Moo96], et les progrès qui seront faits pour lui conférer une résistance aux interférences seront aussi bien des progrès dans la compréhension de nos mécanismes perceptifs.

5. Conclusion

L'analyse de scène est indissociable de la perception en général, et de la parole en particulier. La parole a servi de stimulus privilégié dans la découverte de ses principes, même si elle paraît parfois leur échapper et obéir à des principes qui lui sont propres. Parmi les indices d'organisation importants figure l'harmonicité, lié au voisement, qui permet une amélioration importante de l'intelligibilité. Les systèmes de reconnaissance de la parole auraient bien besoin de profiter eux aussi d'indices de ce type, mais pour l'instant les tentatives de constituer des systèmes d'analyse computationelle (CASA) sont restées à l'état expérimental. Il faut espérer que des progrès tels que la théorie des données manquantes, liés à une meilleure compréhension des mécanismes physiologiques de l'ASA, permettront des avancées dans un avenir proche.

6. Pour en savoir plus

L'ouvrage très complet de Bregman [Bre90] résume la recherche en ASA jusqu'en 1990, et a inspiré l'essentiel de ce qui s'est fait depuis. Des revues plus récentes sont celles de Darwin et Carlyon [DaC95] et Cooke et Ellis [CoE00]. Les articles de Cherry [Che53], Brokx et Nooteboom [Bro82], Warren [War70], Cutting [Cut76], Darwin [Dar81] sont des classiques à lire. Le versant computationnel (CASA) est résumé par Cooke et Ellis [CoE00] (voir aussi [deC00]). La fraîcheur des pionniers est à retrouver dans les thèses de Scheffers [Sch83], Weintraub [Wei85], Cooke [Coo91], Brown [Bro92], Lea [Lea92], Mellinger [Mel91], Ellis [Ell96]. Pour les travaux récents on peut consulter les actes des workshop CASA satellites de l'IJCAI, dont certains ont été publiés [RoO97, SpC99]. Pour les applications de reconnaissance de la parole, et en particulier la théorie des données manquantes, lire Cooke et Green [CoG00], Cooke, Green, Josifovski et Vizinho [CGJ00]. D'autres articles intéressants sont ceux de Lippmann [Lip97, LiC97], Hermansky [Her98]. Pour quelques ressources en ligne, consultez la page: http://www.ircam.fr/pcm/cheveign/sh/casa.html

Bibliographie

[AhT93] Ahmad, S., and Tresp, V. (1993). "Some solutions to the missing feature problem in vision," in "Advances in Neural Information Processing Systems 5," Edited by S. J. Hanson, J. D. Cowan and C. L. Giles, San Mateo, Morgan Kaufmann, 393-400.

[AsS90] Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

[AsS94] Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. 95, 471-484.

[BeM95] Berthommier, F., and Meyer, G. (1995). "Source separation by a functional model of amplitude demodulation.", Proc. ESCA Eurospeech, 135-138.

[Bre90] Bregman, A. S. (1990). "Auditory scene analysis," Cambridge, Mass., MIT Press.

[Bro82] Brokx, J. P. L., and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," Journal of Phonetics 10, 23-36.

[Bro92] Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[BrC93] Brown, G. J., and Cooke, M. P. (1992). "Com-

putational auditory scene analysis: grouping sound sources using common pitch contours," Proc. Inst. of Acoust. 14, 439-446.

[Car94] Carlyon, R. (1994). "Further evidence against an accross-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components," J. Acoust. Soc. Am. 95, 949-961.

[Che53] Cherry, E. C. (1953). "Some experiments on the recognition of speech with one, and with two ears," J. Acoust. Soc. Am. 25, 975-979.

[Coo91] Cooke, M. P. (1991), "Modeling auditory processing and organisation," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[CoB93] Cooke, M. P., and Brown, G. J. (1993). "Computational auditory scene analysis: exploiting principles of perceived continuity," Speech Comm. 13, 391-399.

[CoH00 Cooke, M., and Ellis, D. P. W. (2000). "The auditory organization of speech and other sources in listeners and computational models," Speech Comm., in press.

[CoG00] Cooke, M., and Green, P. (2000). "Auditory organization and speech perception," in "Listening to speech: an auditory perspective," Edited by S. Greenberg and W. Ainsworth, Oxford, Oxford University Press, in press.

[CGJ00] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2000). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Communication (submitted)

[CMG97] Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

[CuD93] Culling, J. F., and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. 93, 3454-3467.

[CuD94] Culling, J. F., and Darwin, C. J. (1994). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. 95, 1559-1569.

[CuS95] Culling, J. F., and Summerfield, Q. (1995). "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay.," J. Acoust. Soc. Am. 98, 785-797.

[Cut76] Cutting, J. E. (1976). "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening," Psychol. Rev. 83, 114-140.

[Dar75] Darwin, C. J. (1975). "On the dynamic use of prosody in speech perception," in "Structure and process in speech perception," Edited by A. Cohen and S. G. Nooteboom.

[Dar81] Darwin, C. J. (1981). "Perceptual grouping of speech components differing in fundamental frequency and onset-time," QJEP A 33, 185-207.

[Dar77] Darwin, C. J., and Bethel-Fox, C. E. (1977). "Pitch continuity and speech source attribution," J. Exp. Psychology: Human Perception and Performance 3, 665-672.

[DaC95] Darwin, C. J., and Carlyon, R. P. (1995). "Auditory grouping," in "Handbook of perception and cognition: Hearing," Edited by B. C. J. Moore, New York, Academic Press, 387-424.

[deC93] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

[deC94] de Cheveigné, A., Kawahara, H., Aikawa, K., and Lea, A. (1994). "Speech separation for speech recognition," Journal de Physique IV 4, C5-545-C5-548.

[deC95] de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

[deC96] de Cheveigné, A., and Marin, C. (1996). "The segregation of frequency-modulated concurrent harmonic sounds," J. Acoust. Soc. Am. 100, 2718.

[deC97a] de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997). "Concurrent vowel identification I: Effects of relative level and F0 difference," J. Acoust. Soc. Am. 101, 2839-2847.

[deC97b] de Cheveigné, A., McAdams, S., and Marin, C. (1997b). "Concurrent vowel identification II: Effects of phase, harmonicity and task," J. Acoust.Soc. Am. 101, 2848-2856.

[deC97c] de Cheveigné, A. (1997). "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

[deC98] de Cheveigné, A. (1998). "Cancellation model of pitch perception," J. Acoust. Soc. Am. 103, 1261-1271.

[deC99a] de Cheveigné, A., and Kawahara, H. (1999).

"Multiple period estimation and pitch perception model," Speech Communication 27, 175-185.

[deC99b] de Cheveigné, A. (1999). "Vowel-specific effects in concurrent vowel identification," J. Acoust. Soc. Am. 106, 327-340.

[deC99c] de Cheveigné, A., and Kawahara, H. (1999). "Missing data model of vowel perception," J. Acoust. Soc. Am. 105, 3497-3508.

[deC99d] de Cheveigné, A. (1999). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. 106, 2959-2972.

[deC99e] de Cheveigné, A. (1999). "Formant bandwidth affects the identification of competing vowels.", Proc. ICPhS, 2093-2096.

[deC99f] de Cheveigné, A. (1999). "Pitch shifts of mistuned partials: a time-domain model," J. Acoust. Soc. Am. 106, 887-897.

[deC00] de Cheveigné, A. (2000). "L'analyse de scènes auditives computationnelle," in "La parole, des modèles cognitifs aux machines communicantes - Développement," Edited by J. Mariani, Paris, Hermès, en préparation, .

[deV99] de Veth, J., Cranen, B., de Wet, F., and Boves, L. (1999). "Acoustic pre-processing for optimal effectivity of missing feature theory.", Proc. Eurospeech, 65-68.

[Dem90] Demany, L., and Semal, C. (1990). "The effect of vibrato on the recognition of masked vowels," Percept. & Psychophys. 48, 436-444.

[Ell96] Ellis, D. (1996), "Prediction-driven computational auditory scene analysis," MIT unpublished doctoral dissertation.

[Ell97] Ellis, D. P. W. (1997). "Computational auditory scene analysis exploiting speech-recognition knowledge," Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio, Mohonk.

[Gre97] Greenberg (1997). "Understanding speech understanding: towards a unified theory of speech perception.", Proc. ESCA Workshop on the auditory basis of speech perception, Keele, 1-8.

[Har96] Hartmann, W. M. (1996). "Pitch, periodicity, and auditory organization," J. Acoust. Soc. Am. 100, 3491-3502.

[Har91] Hartmann, W. M., and Johnson, D. (1991). "Stream segregation and peripheral channeling," Music Perception 9, 155-184.

[Hel77] Helmholtz, H. v. (1877). "On the sensations of tone (English translation A.J. Ellis, 1954)," New York, Dover.

[Her98] Hermansky, H. (1998). "Should recognizers have ears?," Speech Comm. 25, 3-27.

[Lea92] Lea, A. (1992), "Auditory models of vowel perception," Nottingham unpublished doctoral dissertation.

[Lic56] Licklider, J. C. R. (1959). "Three auditory theories," in "Psychology, a study of a science," Edited by S. Koch, New York, McGraw-Hill, I, 41-144.

[Lip97] Lippmann, R. P. (1997). "Speech recognition by machines and humans," Speech Comm. 22, 1-16.

[LiC97] Lippmann, R. P., and Carlson, B. A. (1997). "Using missing feature theory to actively select features for robust speechj recognition with interruptions, filtering, and noise," Proc. ESCA Eurospeech, KN-37-40.

[Lyo94] Lyon, R. (1984). "Computational models of neural auditory processing," Proc. IEEE ICASSP, 36.1.(1-4).

[Lyo83] Lyon, R. F. (1983-1988). "A computational model of binaural localization and separation," in "Natural computation," Edited by W. Richards, Cambridge, Mass, MIT Press, 319-327.

[Mar91] Marin, C. (1991), "Processus de séparation perceptive des sources sonores simultanées," Paris III unpublished doctoral dissertation.

[MaM91] Marin, C., and McAdams, S. (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width," J. Acoust. Soc. Am. 89, 341-351.

[Mar82] Marr, D. (1982). "Representing and computing visual information," in "Artificial Intelligence: an MIT perspective," Edited by P. H. Winston and R. H. Brown, Cambridge, Mass, MIT Press, 2, 17-82.

[Mas90] Massaro, D. W. (1990). "Models of integration given multiple sources of information," Psychological review 97, 225-252.

[McA84] McAdams, S. (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Stanford unpublished doctoral dissertation.

[McA89] McAdams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. 86, 2148-2159.

[MeH92] Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233-245.

[Mel91] Mellinger, D. K. (1991), "Event formation and separation in musical sound," Stanford Center for computer research in music and acoustics unpublished doctoral dissertation.

[Moo96] Moore, R. (1996). "Critique: The potential role of speech production models in automatic speech recognition," J. Acoust. Soc. Am. 99, 1710-1713.

[NOK95] Nakatani, T., Okuno, H. G., and Kawabata, T. (1995). "Residue-driven architecture for computational auditory scene analysis.", Proc. IJCAI, 165-172.

[SpC99] numéro spécial Speech Communication v. 27 nos 3-4, 1999.

[Par76] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911-918.

[Rem94] Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). "On the perceptual organization of speech," Psychological Review 10, 129-156.

[RoO97] Rosenthal, D. F., and Okuno, H. G. (1997). "Computational auditory scene analysis," Lawrence Erlbaum.

[RoD98] Rozogan, A., and Deléglise, P. (1998). "Adaptive fusion of acoustic and visual sources for automatic speech recognition," Speech Comm. 26, 149-161.

[Sch83] Scheffers, M. T. M. (1983), "Sifting vowels," Gröningen unpublished doctoral dissertation.

[Sla95] Slaney, M. (1995). "A critique of pure audition.", Proc. Computational auditory scene analysis workshop, IJCAI, Montreal.

[Sum92] Summerfield, Q. (1992). "Roles of harmonicity and coherent frequency modulation in auditory grouping," in "The auditory processing of speech: from sounds to words," Edited by M. E. H. Schouten, Berlin, Mouton de Gruyter, 157-166.

[SuC92a] Summerfield, Q., and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency," Proc. 124th meeting of the ASA, 2317(A).

[SuC92b] Summerfield, Q., and Culling, J. F. (1992). "Auditory segregation of competing voices: absence of effects of FM or AM coherence," Phil. Trans. R. Soc. Lond. B 336, 357-366.

[Wan95] Wang, A. L.-C. (1995), "Instantaneous and frequency-warped signal processing techniques for auditory source separation," unpublished doctoral dissertation, CCRMA (Stanford University).

[War70] Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," Science 167, 392-393.

[Wei85] Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation," Stanford unpublished doctoral dissertation.

[YDS96] Yost, W. A., Dye, R. H., and Sheft, S. (1996). "A simulated "cocktail party" with up to three sound sources," Perception and Psychophysics 58, 1026-1036.

[ZiW83] Zissmann, M. A., and Weinstein, C. J. (1989). "Speech-state-adaptive simulation of co-channel talker interference suppression.", Proc. IEEE-ICASSP, 361-364.

Faits cosmiques, et faibles masses

Joël Martin

Département d'Astrophysique, de Physique des Particules, de Physique Nucléaire de d'Instrumentation Associée

jmartin@dapnia.cea.fr - http://www-dapnia.cea.fr

UN PETIT MOT DES ORGANISATEURS...

Physicien à Saclay (Commissariat à l'Energie Atomique, CEA) en 1968. Spécialité : recherche fondamentale en physique nucléaire (étude du noyau des atomes), et plus exactement en physique « hadronique » où l'on cherche à « voir » et à comprendre le comportement des particules élémentaires qui s'ébattent à l'intérieur des noyaux (protons et neutrons), et celui des particules encore plus élémentaires (les quarks) que renferment lesdits protons et neutrons. Le noyau atomique est une poupée russe.

Depuis 9 ans, il s'occupe de vulgarisation et de communication au Département d'Astrophysique, de Physique des Particules, de Physique Nucléaire de d'Instrumentation Associée (ouf, fermez le ban). Les intimes disent le DAPNIA. Ce département regroupe 700 personnes au sein de la Direction des Sciences de la Matière au CEA et édite un journal *ScintillationS* qu'anime Joël Martin.

Joël Martin pratique la contrepétrie comme la science, la musique, l'alpinisme et la photo: à fond. Il pratique la contrepèterie pour petits (4 livres) et grands (5 livres). Titulaire de la chaire de Contrepet au *Canard Enchaîné* (rubrique : « Sur l'Album de la Comtesse) depuis 1984. Sujet de la conférence : les fables de la Comtesse, ou comment faire de la contrepèterie avec de la science, et réciproquement.

Voici un petit exercice préparatoire à la conférence : un texte sur les pêcheurs au neutrino (contrepèterie) paru dans Pour la science d'Avril 1999, page 10. Il y a 105 contrepèteries. Bon courage pour arriver au but (contrepèterie) !

ENIGMES

Bien des énigmes de l'Univers mettent simultanément en jeu[1] d'énormes faits cosmiques et de minuscules grains qu'on découvre sans cesse.

Les problèmes de grandeur ont toujours fait bosser les chercheurs. Ils se sont penchés autant sur le bas d'échelles, ce qui implique des tailles très courtes et des faibles masses, que sur les cotes des comètes.

Se prêtant au jeu du néant, un néant fécond, Einstein, grand humaniste couvert d'honneurs dont la description des courbes nécessite beaucoup de veilles, a tordu l'abîme immense. Il a dilaté les temps en les comptant. Mais il n'a pas vu l'extension de l'Univers que des expérimentateurs ont observée en captant des bips, bips qui tombent sur la Terre. Un astronome colérique qui perdait son latin dès qu'il explosait et tenait alors d'incompréhensibles propos sur sa turne, un vrai sabir, a même prétendu qu'un effet Auger des profondeurs zénitales expliquait l'annulation d'une nova ! On a vu une astrophysicienne vexer en parlant des rayons cosmiques sans valeur : on lui avait dit que sa glotte était pleine de muons. En revanche, d'autres chercheurs sensés qui savaient communiquer ont su exploiter ces messages spatiaux, et décoder avec un art consommé, comme le fit tel savant à tête de lion en découvrant ses pulsars.

En face de ces champions des fresques spatiales, chéris des chroniqueurs fréquentant assidûment leurs groupes, on trouve des spécialistes des particules qui testent sans cesse et sans fin. Ces physiciens font une sacrée tête dès qu'ils découvrent de beaux photons. Une spécialiste des quarks a trouvé beaucoup de sites pour la beauté. D'autres expertes maîtrisent parfaitement la réduction des sections efficaces et se montrent indulgentes envers les journalistes qui biffent « proton », ont des attitudes bêtes devant leurs kaons mais fêtent leurs pions. Et ils savent les emmener sur la pistes des chimères.

Des lycéens sortant du bac en se demandant si l'Univers est clos, aux membres de l'Académie fêlés dès qu'ils parlent de dilatation car ils se demandent si l'abîme se tasse, tous les scientifiques supputent les mêmes choses en prenant parfois de drôles de poses comme certains jeunes licenciés qui se font volontiers pompeux et se prennent au jeux des questionnaires. Quelles valeurs pour les spins[2] ? Les champs sont-ils légers ? Un astrophysicien à turban parle souvent de masse cachée.

Les chercheurs se branchent alors sans délai sur de grands mystères sans craindre de vider bien des poches. Il faut des navettes valant beaucoup plus qu'un million et envoyer en l'air des cosmonautes qui se retrouvent décalés dès qu'ils sont en fusées. Mais il faut aussi d'énormes anneaux où ce ne sont pas des vieux bus qui circulent en pure perte car aucun n'attend personne, mais où paressent et se cognent à chaque instant avec une colossale énergie

[1] Le mot jeu est souvent employé par un arrogant spécialiste de Vernes (Vernes Jules) dont les fictions auraient dit-on affermi le règne.

[2] Le grand spécialiste du spin est-il Fermi ?

de drôles de particules par centaines de milliers. Bref, il ne faut pas cesser de payer, ce qui déclenche une avalanche de coûts.

Ainsi, à Kamioka au Japon, les responsables internationaux d'une énorme expérience viennent de provoquer l'effroi des Nippons avec des coûts de taille. Mais cette débauche de béton va susciter beaucoup de thèses comme le montrent l'équipe des chercheurs qui n'ont pas eu le temps de paresser et se font déjà connaître par le bruit de leurs chiffres.

C'est au bout d'une très longue prospection qu'a été lancé le germe des idées tant ces chercheurs étaient en quête de rare. Après bien des passages dans la région de Kamioka, ils ont découvert une immense mine. C'était l'abysse dont ils rêvaient et qui suscita d'emblée de grandes envies de fêtes. Indépendamment du chois des Nippons, ils ont voté avant de se déployer dans des travaux dont ils se disputaient les places. Ils ont commencé par étayer l'abysse car de tels trous ne se comblent pas facilement, puis elle a été complètement plaquée par l'équipe des chercheurs qui rêvaient de murs d'ambre. En définitive, après avoir encollé les murs, ils ont choisi l'habile détecteurs dont ils se sont servie à de multiples exemplaires pour tapisser partout en se contentant de pins, car l'ascèse mais aussi les bains étaient de rigueur[3]. On se souillait facilement avant de pouvoir détecter ce qui était le but que courraient ces traqueurs de l'infime. Infimes ne furent pas les pannes mais ils finirent par arriver au bout et ce ne fut pas en treize ans.

Avoir atteint ce but devant lequel ils ne cafouillèrent jamais leur permet de mieux connaître l'effet des forces faibles. Comment se transforme un neutron qu'on prend sur le nez ? Mais aussi, comment notre Terre peut elle être tapée des milliards de fois par seconde sans qu'on ne sente rien, par d'invisibles grains obus? Les hommes, eh bien, ont mis longtemps à trouver le neutrino. Puis ils l'ont vu de plus en plus léger mais ce n'est pas en le faisant chuter. A-t-il une masse et laquelle, ils en sont perturbés car leurs femmes doutent avec eux. S'il en a une et si elle est trop grosse, ce sera l'effondrement général car le monde pèse trop et finira brisé. S'il n'en a pas ou si elle est toute petite, le monde ne pèse pas assez et finira par s'emballer, tel une énorme bulle qui gonfle et poursuit pour l'éternité son incroyable route.

Avec Kamioka et cette immense mine sans porosité, tout est désormais assez su. Quelques jaloux fustigent la « pègre de physiciens qui exhibent leur neutrinos ». D'après un spécialiste, c'est une vague célébrité du nom d'Aline qui parle ainsi de pègre et pourtant cette personne en a lu dans l'annonce : *« Les neutrinos même maigris émergent massifs du soleil. »*

Le vicomte Aumont

---

[3] On rotait avant de plonger.

Interpréter la prosodie

Albert Di Cristo

Institut de Phonétique, Laboratoire Parole et Langage
Université de Provence (Centre d'Aix)
29, Avenue Robert Schuman
13621 Aix-en-Provence, CEDEX 1, France
albert.dicristo@lpl.univ-aix.fr

ABSTRACT

The central theme of this presentation is the interpretation of prosody. After discussing the general formal and functionals aspects of this problem we present some propositions with the aim of elaborating a general interpretative framework for prosody as well as a system of representation which integrates the different levels of prosodic analysis.

1. INTRODUCTION

« Puisque nous avons certainement une prosodie, on parviendra tôt ou tard à la bien connaître », écrivait l'abbé d'Olivet (1682-1768) dans son *Traité de Prosodie Française*. Trois siècles plus tard, force est de constater que cette prophétie ne s'est que partiellement avérée et que la prosodie demeure, quoi que l'on en dise, un phénomène encore mal connu et plutôt enclin à susciter la perplexité des chercheurs qui s'attachent à poursuivre son étude.

L'intérêt pour la prosodie a pourtant bénéficié, au cours des trente dernières années, d'un essor fulgurant, dont les épisodes les plus marquants concernent, d'une part, son intégration dans le champ de la linguistique formelle (grâce à l'avènement des théories dites « métrique » et « autosegmentale » :[Leb73], [Lib75], [Gol76], [Gol90], [Pie80], [Hay81], [Gol95] et à l'émergence du modèle de la phonologie prosodique [Sel84], [Nes86]. Cf. [Lad96] et [Sha96] pour une vue d'ensemble) et, d'autre part, son impact dans la « mouvance cognitique », notamment dans les secteurs de la psycholinguistique [Bac86], [Cut97] et des neurosciences [Bau97], [Bau00], [Caz00].

Dans le domaine des Sciences du Langage, il se trouve que la recherche prosodique bénéficie également des retombées positives qui sont liées à une évolution significative de la linguistique contemporaine, cette dernière tendant en effet à déplacer son champ d'investigation de l'étude de la langue à celle de ses usages [Fuc92], [Reb98] (voir également le numéro spécial de la Revue *Sciences Humaines* de juin 1995 [Sci95]). De ce fait, la prosodie est appelée à occuper une place de plus en plus importante dans les travaux qui se réclament de la « perspective » pragmatique [Mar87], [Van89], [Cae91]).

On remarquera enfin la participation significative de la France à cette promotion de la prosodie, avec la publication, dans les années 98-99, de quatre ouvrages sur le sujet, signés par des auteurs appartenant à notre communauté scientifique [Hir98], [Mor98], Lac99] et [Ros99]).

L'emprise de la prosodie, qui dépasse largement les frontières des Sciences Humaines, pourrait passer aux yeux de certains pour un fait de mode, une « *prosodimania* », en quelque sorte, s'il n'était établi que la prise en compte à la fois systématique et pluridisciplinaire de cette composante du langage trop longtemps négligée, éclaire d'un jour nouveau quelques paradigmes fondamentaux des recherche sur le langage et la parole, ou suscite leur réinterprétation, voire leur révision. Quelques exemples empruntés à la linguistique, à la psycholinguistique et aux neurosciences suffiront à illustrer ce propos.

En premier lieu, il est établi que les fondements des théories métrique et autosegmentale procèdent d'une rupture épistémologique avec le modèle standard de la phonologie générative [Cho68] que les tenants de ces théories considèrent dorénavant comme un « mode de pensée périmé » [Dur90]. La rupture provient essentiellement du fait que dans le modèle standard les entités prosodiques de la représentation phonologique, telles que les accents et les tons, sont « prisonnières » d'un paradigme de traits hétérogènes, alors qu'elles sont traitées comme des éléments autonomes de la représentation dans les nouvelles approches (d'où le terme consacré « d'auto-segment »). Cette revendication de l'autonomie de la prosodie a, comme on le verra par la suite, des conséquences importantes, à la fois du point de vue de la théorie et de ses applications.

En second lieu, on pourra constater que la prosodie occupe une position de plus en plus centrale dans les courants fondamentaux qui motivent les recherches en psycholinguistique: l'étude de l'acquisition du langage par l'enfant et celle des processus cognitifs et des systèmes de traitement qui participent à l'encodage et à la compréhension de la parole chez l'adulte.

En ce qui concerne le premier thème, il existe de nombreuses évidences empiriques qui attestent le rôle crucial joué par la prosodie dans l'ontogenèse, non

seulement comme véhicule de l'affect, indispensable au maintien des relations interpersonnelles entre l'enfant et ses « interlocuteurs » [Fer89], [Col99]), mais aussi en tant que « dispositif d'amorçage » permettant de stimuler l'acquisition des propriétés phonologiques, lexicales et syntaxiques de la langue maternelle [Deb96], [Vih96], [Nes96], [Jus97], [Kon97]. De ce point de vue, des données récentes sont de nature à relancer le débat relatif au « paradigme insatisfait » (par manque d'évidences expérimentales) de la dichotomie intuitive opposant les langues à « isochronie syllabique » et les langues à « isochronie accentuelle », une question qui a suscité de nombreuses controverses : [Wen82], [Dau83], [Ber89], [Bar94].

Il semblerait en effet [Ram99] que l'on puisse mettre en évidence des indices qui corroborent cette distinction et qui permettraient ainsi de rendre compte des compétences discriminatoires dont font preuve les bébés qui possèdent notamment la capacité de distinguer leur langue maternelle des autres idiomes dès la naissance [Naz98]. Ces performances surprenantes semblent reposer en partie sur une exceptionnelle faculté de résolution temporelle, qui s'exerce dès la naissance [Chr94] et dont les carences sont susceptibles de donner lieu plus tard à des dysfonctionnements langagiers conséquents [Tal93].

Les recherches fécondes entreprises dans la voie que nous venons d'évoquer, outre qu'elles suscitent un questionnement sur l'inventaire des classes rythmiques [Aue93] « ontogénétiquement distinctives » dans le domaine de l'acquisition du langage, se prêtent également à l'évaluation de la valeur opératoire des théories qui occupent l'avant-scène de la linguistique contemporaine, comme la « Théorie des Principes et Paramètres » [Cho81] ou la « Théorie de l'Optimalité » [Pri93].

Une autre illustration édifiante de l'influence de la prosodie sur l'ouverture - ou la mise en cause - de certains paradigmes de recherche en psycholinguistique, concerne la problématique cruciale de l'accès au lexique et de la reconnaissance des mots. Dans cette perspective, on relèvera que, jusqu'à présent, les recherches ont surtout porté sur l'écrit et la reconnaissance visuelle des mots, et que la prosodie a été considérée pendant longtemps comme une variable marginale [Win75], [Seg00].

Bien qu'il soit plausible que la prosodie joue un rôle (qui reste à établir sans conteste) dans la segmentation des énoncés en mots et dans l'activation en ligne des entités lexicales, il peut paraître curieux que les travaux qui s'attachent à modéliser les dispositifs de compétition qui sous-tendent la reconnaissance des mots, qu'il s'agisse du « modèle de la cohorte » [Mar90] ou du modèle de « la densité de voisinage » [Luc90], ne prennent pas en considération le rôle que pourrait jouer l'information prosodique dans ces dispositifs [Lin99]. Quelques recherches isolées montrent cependant que l'information prosodique peut être utilisée dès les stades initiaux de l'activation lexicale, ce qui est vrai notamment des langues à tons, à accentuation mélodique, comme le japonais, ou de certaines langues à accentuation lexicale, comme le néerlandais. Mais, contre toute attente, tel n'est pas le cas de l'anglais sur lequel se sont pourtant centrées la plupart des recherches. Ce qui pourrait alors signifier que ces dernières ont été effectuées sur un idiome qui ne semble pas être représentatif de l'influence exercée par la prosodie sur la reconnaissance des mots [Cut99].

Il ne serait pas surprenant, dans l'optique du thème de recherche que nous venons d'évoquer, que l'on parvienne à montrer que la prosodie joue un rôle non négligeable dans les langues à accentuation fixe (comme le français) qui n'ont pas fait l'objet jusqu'à présent d'une attention aussi soutenue [Ise98].

Construire une cartographie de la neuro-anatomie fonctionnelle de la prosodie dans le cerveau humain en recourant à une interprétation systématique des pathologies prosodiques [Bha94], constitue l'une des tâches les plus préoccupantes et les plus stimulantes de la neurolinguistique actuelle, comme en témoigne la publication en 1999 d'un numéro de la revue *Brain and Language* consacré à cette problématique (une tâche qui pourrait être facilitée par le développement des nouvelles méthodes d'imagerie cérébrale [Bro00]).

Eu égard à la complexité formelle et fonctionnelle de la prosodie (cf. infra), l'enjeu est de taille, dans la mesure où ces recherches contribuent à informer à la fois les modèles d'organisation neurale et les modèles de traitement cognitif.

Si l'on peut montrer, en effet, qu'une opération cognitive particulière impliquant la prosodie est associée à deux zones cérébrales distinctes, on pourra alors en déduire qu'il doit exister une connexion physiologique entre ces zones [Bau00]. Poussée à l'extrême, ce type d'analyse pourrait également avoir des répercussions sur l'évaluation respective des deux modèles concurrents qui ambitionnent d'expliquer le fonctionnement du langage ; le modèle de la « modularité de l'esprit » [Fod83] et le modèle « fonctionnaliste de compétition » [Bat89].

Bien qu'il soit certainement prématuré de répondre à ces questions et en dépit du caractère disparate, voire déroutant des résultats, il apparaît sans conteste que les données dont nous disposons sont à même de réfuter la conception ancienne selon laquelle la prosodie serait placée sous la dominance de l'hémisphère droit. De même qu'elles rendent obsolète l'hypothèse de la spécialisation fonctionnelle [Ros81], d'après laquelle la prosodie émotionnelle serait gérée par l'hémisphère droit et la prosodie linguistique par l'hémisphère gauche.

Les données disponibles sont également très intéressantes, en ce sens qu'elles paraissent mettre en évidence une latéralisation efficiente des paramètres prosodiques, la durée étant gérée principalement par l'hémisphère gauche et la F0 par l'hémisphère droit [Rob90], [Van, 92]. Elles semblent attester, d'autre part, non seulement de l'existence d'un niveau autonome de représentation phonologique de la prosodie, particulièrement évident

dans le cas d'un jargon phonémique [Lou00], mais encore d'une relative autonomie des ordres structurels qui font partie de cette représentation. C'est ainsi, par exemple, que certaines lésions cérébrales peuvent engendrer des dysfonctionnements de l'organisation temporelle des énoncés sans altérer leur organisation tonale [Bou99], ou que les sujets souffrant d'un agrammatisme consécutif à une aphasie de Broca conservent les oppositions de quantité caractéristiques de la prosodie lexicale de leur langue [Nie98].

Mais le plus intéressant concerne les observations qui montrent que la latéralisation est également sensible à la taille et à la complexité des unités qui doivent être traitées dans la parole (ce qui justifie indirectement la distinction théorique entre « prosodie lexicale » et « prosodie supralexicale » que nous proposons d'adopter [Hir98], cf. infra). A partir du moment où l'on constate, à la suite de lésions cérébrales, que la production et la perception des unités prosodiques sont particulièrement sensibles à la tâche et à la variabilité contextuelle, il devient évident que les substrats neuraux de la prosodie sont complexes et qu'ils requièrent probablement l'usage d'une infrastructure cérébrale étendue, impliquant à la fois les régions corticales et souscorticales dans les deux hémisphères [Bal99]. Il en irait donc de la prosodie comme de la syntaxe dont le traitement se fonde sur l'activité « concertée » d'un ensemble de zones cérébrales possédant chacune sa propre spécialisation [Bro00]. Ceci n'est pas étonnant si l'on convient que la prosodie est au coeur même des dispositifs d'intégration supralexicaux de la parole.

Bien évidemment, tous ces résultats ne remettent pas en question l'importance de l'hémisphère gauche dans l'expression et la compréhension du langage, celle-ci constituant certainement un part importante du patrimoine génétique des êtres vivants, dans la mesure où les vocalisations animales semblent être gérées aussi par cet hémisphère [Hau94].

Ce long préambule n'avait pour objectif que de souligner l'importance de la prosodie dans la compréhension du fonctionnement du langage et de la parole. Compte tenu du large consensus interdisciplinaire qui reconnaît aujourd'hui cette importance et de la multiplicité des recherches qui en découle, il devient plus impératif que jamais de s'interroger sur les problèmes que soulève l'interprétation de cette composante centrale de la langue parlée. Dans la première partie de cet exposé, nous passons en revue les principales sources de difficulté de l'analyse prosodique. Nous présentons ensuite un des propositions concernant, d'une part, l'élaboration d'un cadre interprétatif de la prosodie et, d'autre part, une définition explicite de ses niveaux d'analyse et de représentation.

2.DES DIFFICULTES DE L'ANALYSE PROSODIQUE

2.1 Définir la prosodie

Il ne serait pas superflu que toute recherche prosodique procède d'une définition explicite de la prosodie. Cette contrainte aurait au moins l'avantage d'instaurer une meilleure compréhension entre les chercheurs qui entreprennent son étude sous des angles parfois fort différents. Peut-être serait-on également à même, en procédant de la sorte, de réviser les définitions réductrices ou opaques qui entachent les dictionnaires et les encyclopédies « de tout poil » (la comparaison de ces définitions constituera, pour qui veut bien la réaliser, une expérience particulièrement édifiante).

Définir la prosodie (comme c'est souvent le cas) par référence à l'étude des paramètres physiques que sont la F0, la durée et l'intensité, par exemple, conduit à qualifier la prosodie comme un fait de substance. Cette définition est manifestement réductrice, voire biaisée, car elle ne rend justice ni à la fonctionnalité des faits prosodiques, ni à leur organisation systémique. Nous proposons ci-après une définition personnelle de la prosodie qui ne prétend pas être définitive. Nous lui adjoignons deux définitions complémentaires (l'une d'entre elles étant appliquée au terme intonation, souvent employé indifféremment avec celui de prosodie) qui nous semblent particulièrement représentatives de la conception actuelle de la prosodie.

- « *La prosodie (ou la prosodologie) est une branche de la linguistique consacrée à la description (aspect phonétique) et à la représentation formelle (aspect phonologique) des éléments de l'expression orale tels que les accents, les tons, l'intonation et la quantité, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F0), de la durée et de l'intensité (paramètres prosodiques physiques), ces variations étant perçues par l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation pragmatique dans le flux du discours.* »
- La prosodie est « *une structure grammaticale possédant une organisation qui lui est propre.*» [Be96].
- Le terme intonation « *fait référence à l'usage qui est fait des traits phonétiques suprasegmentaux pour véhiculer, au niveau 'postlexical' ou de la phrase, des signifiés pragmatiques d'une façon linguistiquement structurée.* » [Lad96].

Comme on l'aura constaté, nous n'utilisons pas le terme « suprasegmental », hérité de la tradition américaine [Leh70], en raison de son ambivalence notoire [Ros99], étant donné que dans la littérature il fait référence soit à des entités plus larges que les phonèmes (ce qui est

discutable), soit à des faits non segmentables (ce qui n'est pas tenable), soit encore à des couches d'éléments phoniques superposables à la ligne phonémique. Seule cette dernière acception est conciliable avec l'approche multilinéaire de la prosodie, inspirée du « modèle de la partition orchestrale » de Hockett, qui prévaut de nos jours. Mais comme l'intonation est décrite et représentée dans cette approche en termes de segments tonals, il serait pour le moins curieux, voire confondant, de parler alors de « segments suprasegmentaux ».

.Pourquoi la prosodie est-elle si difficile à interpréter ? Afin de fournir des éléments de réponse à cette question qui nous paraît cruciale, nous proposons d'examiner un certain nombre de points relatifs à l'essence et à la fonctionnalité des éléments prosodiques.

Le terme « essence » fera référence ici aux propriétés formelles des systèmes prosodiques, à la spécificité des représentations phonologiques qui sous-tendent ces systèmes [Hir83], [Hir87] et à la nature de leur matérialité. Dès à présent, nous mettrons en avant l'idée que la prosodie ne se résout pas à des phénomènes de surface, comme pourraient le laisser croire de nombreux travaux, mais qu'elle fait partie des représentations mentales de la langue et des cadres de référence qui les constituent. Ce qui signifie qu'il existe bien une prosodie linguistique indépendante des canaux qui servent à la véhiculer, comme en témoigne notamment « l'existence » d'une prosodie de la langue des signes chez les malentendants [San99].

2.2 Spécificité formelle de la prosodie

Du point de vue formel, la prosodie peut être considérée comme un système complexe (ou comme un ensemble de systèmes) faisant partie intégrante de ce « système de systèmes » qu'est la langue. Sa complexité provient d'une part du fait (et il faut y voir ici un paradoxe) qu'elle constitue une structure autonome reliée aux autres systèmes de la langue comme la morphologie, la syntaxe et la sémantique et, d'autre part, du fait qu'elle se présente elle-même comme un « supra-système » composé de trois ordres structurels relativement indépendants (ce qui justifie en partie les nombreuses tentatives de modélisation « autonomistes »), mais néanmoins interactifs : l'ordre de structuration métrique, l'ordre de structuration tonale (ou mélodique) et l'ordre de structuration temporelle. Ces ordres structurels s'appliquent au lexique (et on parlera alors de « prosodie lexicale ») et/ou à des unités de rang supérieur (et il sera question dans ce cas de « prosodie supralexicale »). Appliqués au lexique, les trois ordres de structuration sont gérés respectivement par des oppositions d'accents, de tons et de quantité. Appliqués à des unités de rang supérieur, ils concernent respectivement la gestion du rythme, de l'intonation et du tempo des énoncés et du discours.

De nombreuses approches récentes (cf. [Dic99] pour une discussion sur ce sujet) admettent de façon plus ou moins explicite la primauté de l'organisation métrique sur l'organisation tonale, aussi bien dans la parole que dans la musique [Dra98]. Cette organisation métrique repose sur la mise en œuvre de deux dispositifs fondamentaux correspondant d'une part à la distribution et à l'extraction (si on se place du point de vue de la perception) des battements, c'est-à-dire des proéminences associées aux unités métriquement fortes et, d'autre part, à la segmentation du continuum de parole en mesures, ces dernières formant le squelette métrique auquel s'appliquent les modulations constitutives de la mélodie (qui est la substance auditive de l'intonation), ainsi que les marques locales (allongements et pauses) et globales (tempo) de l'organisation temporelle [Due87]. Les observations précédentes appellent plusieurs commentaires sur la nature des interactions entre ces paramètres prosodiques abstraits et les choix théoriques et méthodologiques qu'elles suscitent

En premier lieu, on remarquera que les deux entités de l'organisation métrique que sont les proéminences et les mesures font l'objet, dans les théories métriques actuelles les plus influentes, d'une « représentation conjuguée » [Hal87, 95]. Dans ce type de représentation, qui utilise le formalisme de la « grille métrique étiquetée », l'attribution d'une proéminence à une unité de la chaîne sonore donne lieu à la formation d'une mesure correspondant à un domaine particulier, comme le pied métrique, le syntagme accentuel, etc. C'est ainsi que la construction d'un domaine donné (comme , par exemple, le pied métrique) , que signale un parenthésage, effectué au niveau de la ligne (l) de la grille métrique, va entraîner la projection de la tête de ce domaine sur la ligne supérieurs ($l+1$). Les têtes figurant sur cette ligne seront regroupées au sein d'un nouveau domaine qui donnera lieu à son tour à la projection de sa tête sur la ligne (l+2) et ainsi de suite, jusqu'à la projection ultime de la tête représentant le niveau de proéminence le plus élevé de la séquence concernée. Dans les approches qui ne reconnaissent pas le bien-fondé de cette représentation conjuguée, la structure métrique est formée de séquences d'auto-segments correspondants aux proéminences et la grille métrique peut être construite sans faire référence à la constituance. [Lak93]

La primauté de l'organisation métrique s'exprime, dans les modèles formels qui adhèrent à la théorie « métrique-autosegmentale », par le choix d'une option selon laquelle les segments tonals constitutifs de l'intonation sont rattachés aux accents et aux frontières des constituants prosodiques [Pier80], [Hir84]. Dans cette perspective, les proéminences accentuelles, qui sont les points d'ancrage de la structure prosodique, assument un double rôle. En effet, dans la mesure où la proéminence est signalée par une variation mélodique (cas général du « pitch accent », conformément à la théorie de [Bol58] généralement acceptée), ce changement mélodique signale effectivement la présence d'un accent et participe, de surcroît (par la direction de son mouvement), à la

construction de la ligne mélodico-rythmique de l'énoncé que de nombreux auteurs qualifient d'intonation.

La démarche que nous venons d'exposer est mise en œuvre dans le module prosodique du système de synthèse du français à partir du texte SYNTAIX que nous avons développé récemment à l'Institut de Phonétique d'Aix [Dic97]. Dans ce module, les segments tonals (tons L, H, etc) constitutifs de l'intonation sont en effet projetés à partir d'une représentation métrique des énoncés, cette dernière étant établie sur la base des informations fournies par le parseur du module TAL qui opère via l'application d'un algorithme de type « chink and chunk » [Lib92]. L'un des avantages de cette approche est qu'elle permet de conserver une représentation symbolique de la prosodie jusqu'aux stades ultimes de sa dérivation. Ce qui se traduit notamment par une économie drastique du nombre des règles prosodiques qui n'excède pas la dizaine dans SYNTAIX.

Dans les paragraphes qui précèdent, nous avons fait allusion à la notion de constituance qui est au cœur même des problématiques de l'analyse prosodique. Le modèle de la phonologie prosodique : [Sel78], [Sel84], [Nes86]], [Del95], [Truc95], [Sha96] a pour axiome principal que la prosodie est une structure constituante autonome et que les constituants prosodiques qui participent à la construction de cette structure en forment les primitives phonologiques. Bien que la réalité des unités prosodiques ne puisse être contestée, le dogmatisme du modèle cité en référence ne manque pas de susciter des interrogations et des controverses.

Dans sa version standard, le modèle postule les niveaux de constituance suivants (du rang le plus élevé au rang le plus bas) : l'Enoncé (*Phonological Utterance*), le Syntagme Intonatif (*Intonational Phrase*), le Syntagme Phonologique (*Phonological Phrase*), le Groupe Clitique (*Clitic Group*), Le Mot Prosodique (*Prosodic Word*), le Pied (*Foot*), la Syllabe (*Syllable)* et optionnellement la More (*Mora*). Tous les auteurs qui travaillent dans la mouvance de la phonologie prosodique n'ont pas recours à tous ces niveaux, ce qui est logique si l'on admet que certains constituants peuvent être attestés dans une langue et pas dans une autre. Les problèmes émergent surtout du fait que ces auteurs utilisent souvent un même terme pour désigner des groupements différents ou qualifient un groupement donné par des termes différents. Par exemple, Le Mot Prosodique de [Sel78] correspond à la fois au Groupe Clitique et au Mot Prosodique de [Nes86]. De même, le Syntagme Phonologique de [Nes 86] se voit qualifié de Syntagme Majeur, ou de Syntagme Mineur par [Sel84) et de Syntagme Intermédiaire ou de Syntagme Accentuel par [Bec86]. Il est clair que ces divergences confondantes sont l'émanation des nombreuses incertitudes qui planent encore sur le statut des unités prosodiques. A ce propos, deux thèmes de discussion nous paraissent particulièrement cruciaux.

D'une part, il est affirmé que les constituants prosodiques sont principalement des « domaines d'application de règles phonologiques » et que ces règles président à leur fondement. A titre d'exemple, le Syntagme Phonologique est défini par référence : a) à la règle de liaison : la liaison est possible entre des items lexicaux appartenant à un même Syntagme mais pas entre des items consécutifs relevant de deux Syntagmes différents (ce qui est erroné) et b) à la règle « d'évitement de la collision accentuelle » qui stipule que deux syllabes accentogènes adjacentes ne peuvent être accentuées qui si elles sont séparées par la frontière d'un Syntagme Prosodique (ce qui semble être attesté). Mais en même temps, ce constituant est défini par référence à la syntaxe, dans la mesure où les règles qui décrivent sa formation sont fondées sur des principes de projection propres à la syntaxe X-barre [Pos00]. On peut s'interroger, dans ces conditions, sur les limites réelles de cette autonomie prosodique que revendique le modèle standard. Nous estimons, en effet, que l'existence d'un constituant prosodique doit être étayée en ayant recours non pas à des critères syntaxiques ou sémantiques, mais à des critères spécifiquement prosodiques, qui relèvent précisément des trois ordres structurels que nous avons déjà évoqués. Par exemple, l'Unité Intonative (ou le Syntagme Intonatif), dont l'existence fait l'objet d'un consensus notable, devrait être identifiée formellement et phonétiquement, comme nous l'avons proposé [Hir98], en termes de caractéristiques globales, itératives et locales, les deux premières ayant trait au signalement de la cohésion interne de l'unité et les dernières à celui de ses limites (On se reportera, pour une illustration de ces propos, à [Dic93,96]).

D'autre part, le modèle standard suppose que la constituance prosodique est contrainte par une hiérarchie stricte. Cette hypothèse, qualifiée dans la littérature de « *Strict Layering Hypothesis* » prédit qu'un constituant de rang (n) ne peut que dominer des constituants de rang (n-1) et que ces derniers doivent être entièrement compris dans (n). Bien que depuis les travaux de [Mar72], on relève un large consensus sur la nature hiérarchique de la constituance prosodique, plusieurs chercheurs réfutent le caractère strict de cette dernière [Bec86], [Lad86], [Dic96], dans la mesure où il a été montré qu'un constituant de rang déterminé peut dominer directement des constituants qui ont situés à deux ou trois niveaux inférieurs de la hiérarchie, ou qu'un constituant donné peut dominer directement un constituant de même rang, ce qui permet, par voie de conséquence, l'expression de la récursivité dans la structure prosodique (ce qui est le cas, notamment, lorsqu'un Syntagme Intonatif est emboîté dans un autre Syntagme Intonatif). L'analyse de la parole spontanée permet de mettre en évidence de nombreuses déviations par rapport à l'hypothèse de la hiérarchie stricte. Il semble en effet que, dans ce style de parole, la formation de la constituance ne procède pas de l'application d'un cadre préétabli, mais d'une « construction en ligne », motivée essentiellement par des contraintes discursives [Dic99], [Del00]

Outre la problématique de la multilinéarité structurelle des systèmes prosodiques et les interrogations que soulève le

concept de constituance, la question des primitives formelles de l'analyse prosodique se pose également avec acuité. En ce qui concerne plus particulièrement l'intonation, deux approches s'affrontent sur ce terrain, que nous qualifierons respectivement d'approche « holistique » et d'approche « autosegmentale » (ce qui, du reste, ne manque pas de rappeler le vieil antagonisme opposant les partisans de l'analyse prosodique en termes de configurations et les défenseurs de l'analyse en termes de niveaux [Bol 51]. Dans l'approche holistique l'intonation est décrite sous la forme de patrons mélodiques prototypiques [Vai74, 75], de configurations globales constitutives d'un lexique intonatif [Au91], ou encore de contours stylisés, considérés comme étant représentatifs des différents « intonèmes » (ou des morphèmes intonatifs) fonctionnellement distinctifs dans une langue donnée [Del66]. Il arrive que l'approche holistique s'inscrive dans la perspective superpositionnelle [Grø95], [Möb95], suivant laquelle la construction des contours résulte précisément d'une superposition de domaines, comme, par exemple, la composante accentuelle et la composante syntagmatique dans le modèle largement répandu de [Fuj79]. Dans ce cas, les primitives de la description sont les composantes superposables et non le contour global. Par ailleurs, certaines approches linguistiques, qui présentent une parenté évidente avec la mouvance holistique, procèdent d'une conception hiérarchique, dans la mesure ou l'organisation des contours est régie par des principes de dominance. On citera plus particulièrement ici le modèle morphologique de [Ros99] et le modèle phonologique de [Mar87].

L'approche autosegmentale se distingue de la précédente par le fait qu'elle décrit l'intonation en termes de séquences de segments tonals engendrées au moyen d'une grammaire à états finis. C'est en ce sens qu'elle est qualifiée de « linéaire » : Les chercheurs qui se rallient à cette approche adoptent en principe la théorie dite « des deux niveaux » (représentés par les tons L et H), inspirée de l'analyse tonale des langues africaines. Ils divergent cependant sur l'identité des éléments de la structure auxquels sont rattachés les segments tonals dans la représentation. C'est ainsi que les tons sont associés aux accents et aux frontières dans le modèle linéaire standard de [Pie80], aux constituants prosodiques, dans le modèle linéaire-hiérarchique de [Hir84], et aux syllabes, dans le modèle intonosyntaxique de [Mer87]. De ce point de vue, la problématique de l'alignement des segments tonals avec, d'une part, les autres éléments de la structure prosodique et, d'autre part, les unités constitutives de la chaîne segmentale, constitue une des pistes de recherche les plus prometteuses, aussi bien pour parfaire notre connaissance des dispositifs structurels de la parole que pour des applications aux technologies de la parole.

Recenser les mérites respectifs des approches holistiques et autosegmentales nous entraînerait trop loin, mais on pourra se reporter à [Lad95, 96] pour une discussion approfondie sur ce sujet. On s'attardera cependant à présenter quelques arguments qui plaident en faveur de l'approche autosegmentale.

En premier lieu, l'approche autosegmentale utilise le même formalisme pour décrire les tons et l'intonation, ce qui évite d'avoir recours à une formalisation ad hoc pour décrire cette dernière.

En second lieu, l'adhésion générale à la « théorie des deux niveaux » facilite l'élaboration d'outils de transcription et d'étiquetage simples et fiables qui s'avèrent fort utiles pour les comparaisons inter-langues. Actuellement, il existe trois systèmes de ce type en usage : ToBI [Sil92], INTSINT [Hir98, 00] et IViE [Gra98]. Nous ne pouvons ici, faute de place, nous étendre sur les particularités de ces systèmes. Nous noterons cependant que l'application de ToBI à une langue autre que l'anglo-américain peut faire problème, dans la mesure où ce système présuppose une connaissance approfondie de la phonologie prosodique de la langue. Ce n'est pas le cas des deux autres systèmes, qui présentent un outil de transcription de l'intonation relativement neutre vis à vis de la langue, qu'il s'agisse d'une transcription phonétique, comme dans IViE ou d'une transcription phonologique de surface, comme dans INTSINT (cf. la dernière section pour une illustration).

Enfin, il est établi qu'une approche holistique, de même qu'une approche superpositionnelle, ne sont pas en mesure de rendre compte de certains phénomènes importants pour la représentation des structures intonatives [Bru77], [Lad95], [Pos00]. Il s'agit notamment de la spécification des associations entre la structure intonative et le texte (c'est-à-dire, des régularités d'alignement), et des effets de dynamique tonale qui participent à la hiérarchisation intonative en s'appliquant à des accents individuels et non la totalité d'un domaine. Selon [Lad95] ce fait, parmi d'autres, ne plaide ni en faveur d'un modèle superpositionnel, ni en faveur d'un modèle linéaire radical, comme celui de [Pier80], mais plutôt en faveur d'un modèle linéaire qui incorpore une structure métrique indiquant les niveaux de constituance. En d'autres termes, un modèle linéaire hiérarchique.

Ces dernières remarques nous conduisent à élargir la problématique et à affirmer qu'il ne suffit pas de spécifier les séquences de tons et la constituance prosodique pour donner une description phonologique adéquate de l'intonation (i.e. qui rend compte de toutes les régularités formelles) Il importe également de spécifier les distinctions phonologiques de registres et l'identité des facteurs linguistiques qui les motivent. Nous abordons ici une dimension qui a été peu explorée dans les recherches intonologiques, celle des distinctions intonatives qui sont « orthogonales » par rapport à la séquence des segments tonals et qui ont souvent été décrites, de ce fait, comme des distinctions graduelles [Lad90]. A ce propos, nous devons évoquer un autre problème difficile qui entrave l'interprétation de la prosodie et que nous exposerons succinctement. Les fonctions signifiantes de la prosodie s'exercent dans une langue par la mise en oeuvre

conjointe de modes d'expression discrets (ou catégoriels) et non-discrets (ou graduels). Les premiers relèvent de la description phonologique à proprement parler et les seconds de la description phonétique. Il se trouve que la distinction entre ces deux modes est plus difficile à interpréter pour la prosodie que pour tout autre système linguistique et que les descriptions prosodiques les confondent souvent. Il est rare en effet qu'une description prosodique distingue explicitement une phonologie d'une phonétique prosodique Les difficultés proviennent en partie du fait que la « gradualité », qui opère a la fois sur le plan paradigmatique (variabilité de la dynamique tonale) et syntagmatique (variabilité de l'alignement), est une composante permanente et nécessaire de la prosodie où elle assume, notamment, des fonctions de quantification diversifiées, semblables à celles que remplissent les quantificateurs verbaux (cf. Mox93] pour une discussion générale sur le sujet). D'autre part, l'hypothèse séduisante selon laquelle les variations graduelles occuperaient l'espace tonal laissé disponible par les distinctions prosodiques phonologiques (Gus99), implique, bien évidemment, que l'on soit en mesure d'identifier ces dernières. Ce qui n'est pas chose aisée, surtout si l'on est amené à constater que l'appareil des tests de perception catégorielle ne semble pas constituer un outil adapté à la prosodie [Pos00]. Il est donc urgent, dans cette perspective de travailler à l'élaboration d'une méthodologie adéquate.

La complémentarité des oppositions phonologiques et des distinctions graduelles est en corrélation avec la variabilité prosodique, notamment avec la variabilité prosodique situationnelle, qui a donné lieu à de nombreuses recherches sur les propriétés des styles prosodiques. Plusieurs de ces travaux s'intéressent plus particulièrement à l'étude des caractéristiques de la prosodie de la parole dite « spontanée » que l'on compare généralement à la lecture [Es88],[Bla95]. Ces études s'appuient souvent sur l'hypothèse plus ou moins explicite que le passage d'un style à l'autre déclencherait un processus de « *code switching* ». Nous n'adhérons pas à cette idée et nous pensons, en revanche, que tous les styles font usage d'un même « système prosodique noyau », spécifique à langue, et que la variation stylistique est de type paramétrique, ce que semblent du reste confirmer des recherches récentes sur le français [Ast99], [Pos00]. La modélisation de la variabilité prosodique stylistique, qui constitue une étape importante des recherches en prosodie, en vue notamment d'une application à la synthèse de la parole, devra donc s'attacher à mettre en évidence l'ensemble des paramètres qui concourent à l'identification des styles, étant entendu que ces derniers sont parfaitement identifiables, même en l'absence d'information verbale [Léo93].

Sans abandonner la perspective formelle qui a constitué jusqu'à présent le thème central de cette seconde partie, nous souhaitons aborder maintenant un des problèmes les plus épineux de l'analyse prosodique : celui des ruptures de correspondances entre les niveaux d'interprétation.

Si nous considérons, par exemple, les variations de la Fréquence fondamentale (F0) qui constitue la substance physique de l'intonation, il s'avère qu'il n'y a pas de relation terme à terme entre ces variations physiques et les significations linguistiques et para-linguistiques qu'elles sont censées véhiculer. Il y a plusieurs raisons à cela.

La première est que, contrairement à ce que pratiquent de nombreuses études, il n'est pas loisible d'établir des relations directes entre ces variations physiques (fusent-elles plus ou moins stylisées) et la signification des énoncés. Cette mise en correspondance entre expression et contenu ne peut être effective que si l'on prend également en considération les niveaux de représentation phonologique et phonétique de la prosodie (On se reportera à [Lad93] qui a développé une argumentation convaincante sur ce sujet. (cf. aussi la dernière partie de cet exposé).

La seconde raison est que les courbes de F0 actualisent la conjonction d'un ensemble de contraintes (qu'il serait trop long d'énumérer ici), parmi lesquelles figurent, outre celles qui relèvent de la projection des instructions linguistiques intentionnelles du locuteur, celles qui sont régies par les propriétés dynamiques des systèmes de production de la parole, comme, par exemple, les variations microprosodiques intrinsèques (qui dépendent de la nature des segments phonémiques), les variations co-intrinsèques (qui dépendent de la coarticulation [Dic86]) et la déclinaison (qui dépend en partie du contrôle exercé par le système respiratoire) [Vais83]. L'identification et la neutralisation de ces contraintes constitue donc une condition préalable à l'interprétation « linguistique » des tracés de F0 [Dic85].

La troisième raison est que les ruptures de correspondance se manifestent déjà au niveau de la substance, dans la mesure où il n'y a pas de relation terme à terme entre les variations de F0 et la mélodie. Certaines variations de F0 perceptibles (les variations microprosodiques) ne sont pas prises en compte dans la perception de la ligne mélodico-rythmique, alors qu'elles peuvent avoir valeur d'indice au plan segmental. De même, une variation importante de F0 peut n'avoir qu'une faible incidence sur la perception de la mélodie et inversement, ce qui pourrait nous inciter à pencher en faveur d'une « *théorie quantale* » de la prosodie, cette dernière étant d'ailleurs tout-à-fait compatible avec une modélisation en termes de points-clés ou de points-cibles qu'adoptent de nombreux auteurs.

Pour résumer d'une phrase les divers problèmes que nous venons d'évoquer nous dirons que toutes les différences prosodiques mesurables ne sont pas perceptibles et que toutes les différences perceptibles ne sont pas nécessairement perçues [Har90]; quand elles le sont, elles ne sont pas obligatoirement signifiantes et quand elles sont signifiantes, elles ne sont pas forcément porteuses de distinctions phonologiques.

Il resterait un dernier point à envisager, en ce qui concerne la matérialité des systèmes prosodiques et les

difficultés d'interprétation qu'elle suscite. Il s'agit du caractère pluriparamétrique de la prosodie et des effets qu'engendre l'interaction des divers paramètres vis à vis de la perception de la mélodie, des proéminences et des niveaux de frontière. Mais, faute de place, nous ne traiterons pas ici de cette question qui a fait l'objet d'importants développements dans [Ros81].

2.3 Pluralité fonctionnelle de la prosodie

L'ambition d'interpréter la prosodie ne devrait pas se donner pour seuls objectifs de mettre en lumière son architecture formelle (démarche de la phonologie autonomiste) et/ou de décrire les variations paramétriques qui actualisent cette architecture (démarche de la phonétique traditionnelle). Procéder de la sorte revient en fait à manipuler des objets linguistiques, sans se préoccuper de leurs usages ni de leurs fonctions, ce qui est à vrai dire assez représentatif d'une certaine mouvance de la linguistique contemporaine qui tend à marginaliser toute considération sémantique, pragmatique ou fonctionnelle [Bou99]. vérité, la problématique de l'interprétation de la prosodie est éminemment concernée par l'élucidation de ses fonctions signifiantes qui se caractérisent, il faut bien l'admettre, par une singulière hétérogénéité. L'hétérogénéité fonctionnelle de la prosodie provient en partie du fait que les signes qu'elle véhicule occupent la totalité du champ sémiotique : du symbole au symptôme, en passant par l'icône. Elle tient aussi au fait qu'un même signal prosodique peut revêtir, selon le contexte de l'échange verbal, des significations diverses (ce qui évoque l'idée d'une sémiotique « *à géométrie variable* ». Enfin, cette hétérogénéité se manifeste par le fait qu'un même signal prosodique à le pouvoir d'exprimer simultanément des informations de nature très différentes. On pourrait parler à cet égard d'une sémiotique « *syncrétique* » dont seul un modèle de représentation fonctionnelle multilinéaire serait peut être en mesure de rendre compte. Quelles sont les fonctions de la prosodie dans l'exercice du langage et dans la communication interpersonnelle ?

La fonction générale de la prosodie est avant tout une fonction d'assistance à l'encodage et au décodage de la parole. Il est admis en effet que le Syntagme Phonologique [Whe97] et l'Unité Intonative [Lav91], [Lev89] constituent des unités de planification de la parole. Il est clair que cette planification prosodique a des incidences notables sur la transformation des séquences de représentations lexicales discrètes en énoncés rythmiquement bien formés (ce dont doit tenir compte un modèle de production de la parole) et sur la réalisation des segments phonémiques [Fou99] (ce dont devrait peut être tenir compte la synthèse de la parole) L'émergence précoce des Unités Intonatives, probablement au cours de la préparation conceptuelle des messages [Ind00], où elle semble être intimement liée à la planification de la gestuelle para-verbale [Gua91], [Mne92], est mise en évidence par l'étude des lapsus, notamment des lapsus par anticipation qui montrent que les structures prosodiques et syntaxiques sont construites avant la planification lexicale [Ros98]. Sur l'autre versant de la « chaîne de communication verbale », il est également patent que la « continuité prosodique », telle que l'a définie et analysée [Dar75], apporte une contribution significative à l'intelligibilité et à la perception de la parole. On peut également s'interroger sur le rôle que pourrait jouer la « prosodie silencieuse » dans l'écriture ou sa fonction mnémonique dans la prise de notes à l'écrit ou lors d'une conférence par exemple

Outre ces fonctions générales d'assistance à la production et à la perception de la parole et du langage, la prosodie assume un ensemble de fonctions linguistiques, para-linguistiques et extra-linguistiques qui consistent à structurer la langue et le discours, à contextualiser les énoncés et leurs auteurs, à objectiver les modalités illocutoires, à réguler les interactions verbales, à exprimer l'affect et à caractériser le sujet parlant ainsi que le style discursif qu'il adopte.

La fonction structurale de la prosodie occupe une place centrale dans le langage, dans la mesure où elle agit comme un principe organisateur qui façonne le matériau verbal et met en perspective l'information. Cette fonction « grammaticale » - qui s'exerce via la conjonction de trois dispositifs tels que le jeu des proéminences, la démarcation et le liage - s'applique à tous les niveaux de la chaîne linguistique, depuis l'unité lexicale (on se référera ici au rôle déterminant que joue la prosodie dans l'identification des morphèmes qui forment le mot [Gar65]) jusqu'au texte constitutif du discours [Swe93].

Avec la problématique du « parsing prosodique », nous touchons à une des questions les plus débattues de la prosodologie. Le « *parseur* » prosodique opère-t-il *per se*, comme le prétendent certains partisans de la phonologie prosodique (cf. ci-dessus, la définition de Beckman) ? ou bien est-il assujetti aux contraintes que lui impose la syntaxe et/ou la sémantique ? Bien qu'il existe suffisamment de preuves empiriques pour réfuter la conception de l'isomorphisme de la prosodie et de la syntaxe [Mon93] (encore, faudrait - il préciser : quelle prosodie et quelle syntaxe !), il n'est pas possible, en l'état des connaissances, d'apporter une réponse définitive à cette question. Il est clair cependant que l'aptitude des locuteurs à désambiguïser certaines formes d'ambiguïtés syntaxiques au moyen de la prosodie (mais encore faut-il distinguer entre ambiguïtés « de salon », dont sont friands les linguistes et ambiguïtés avérées !) dénote que la syntaxe peut imposer des contraintes à la prosodie.[Sha98]. L'interrelation syntaxe/prosodie est probablement prégnante dans l'ontogenèse. Pinker [Pin93] cite plusieurs travaux qui montrent que les bébés préfèrent, au stade pré-linguistique, les énoncés qui manifestent une congruence entre intonation et syntaxe à ceux qui exhibent une violation de cette congruence. Les Syntagmes Phonologiques, qui représentent à la fois des entités syntaxiques et prosodiques, sont des objets linguistiques accessibles aux bébés grâce à l'usage systématique de marqueurs prosodiques ostensibles dans

la parole maternelle adressée (ou « *motherese* ») [Meh00], [Col99]. Pour en revenir à l'adulte, il existe plusieurs façons d'évaluer la nature des relations entre intonation et syntaxe qui relèvent toutes, en définitive, de la même problématique : connaître comment le sujet (en l'occurrence, l'auditeur) établit la structure syntaxique des énoncés. Diverses pistes se présentent alors. La première concerne l'hypothèse du rôle de la prosodie dans le traitement syntaxique précoce et « en ligne » des énoncés. Cet aspect est lié, d'une part, à la résolution temporaire des ambiguïtés potentielles, notamment des « ambiguïtés d'attachement » [Mar 92], et d'autre part, à l'incidence de l'information prosodique sur le traitement syntaxique subséquent de l'énoncé en cours de réalisation, ce qui revient à tester l'hypothèse de la « fonction prédictive » de la prosodie [Fón79]. Des investigations récentes, qui ont recours à la méthode des « potentiels évoqués », sont de nature à valider ces hypothèses, car elles montrent notamment qu'une prosodie incongrue peut avoir des effets inhibants sur l'étape initiale du parsing syntaxique [Jes98].

Une autre piste intéressante sur la nature des liens entre la prosodie et la syntaxe concerne l'estimation du rôle de la prosodie dans le traitement des relations de dépendances syntaxiques discontinues [Swi98], qui sont d'une fréquence élevée en français parlé [Bla90], [Gül95]. Cette dernière remarque nous amène à évoquer une troisième piste encore peu explorée, qui concerne cette fois l'étude des rapports entre la prosodie et la syntaxe des énonciations, c'est-à-dire la macro-syntaxe, qui dépasse largement le cadre de la phrase sur laquelle ont porté jusqu'à présent la plupart des investigations [Ber90], [Sab96]. L'élargissement du champ de la recherche conduit cependant à s'interroger dès à présent sur la valeur fonctionnelle de certaines macro-unités prosodiques, comme le paraton [Yul80], la période [Haz83] et sur les diverses stratégies prosodiques qui participent à l'organisation discursive [Swe93]. Mais ces investigations nécessitent le support d'un modèle de discours abouti qui soit compatible avec l'analyse prosodique [Gro92], [Ber93], [Cah97], [Reb98].

La nature exacte des relations entre prosodie et syntaxe, qui demeure mal connue [Seg00], est en partie occultée par l'exercice des contraintes métriques (ou rythmiques) et des contraintes sémantico-pragmatiques qui motivent également l'organisation prosodique. Il ne faut pas perdre de vue, en effet, que l'un des rôles majeurs de la prosodie est « d'empaqueter » l'information [Val92] et de construire « l'échelle focale » des énoncés, selon la signification que nous attribuons à cette expression [Dic99]. Bien que la prosodie active « en ligne » la compréhension des énoncés, il est possible que le « phrasé » prosodique n'influence pas directement les décisions syntaxiques, mais qu'il opère essentiellement au niveau de l'interprétation sémantique et pragmatique des énoncés [Pyn99]. Le problème de la « construction » de la syntaxe par l'auditeur représente encore un mystère, il reste à expliciter dans quelle mesure le calcul qui préside à cette construction est réceptif à diverses sources d'informations :sémantique, pragmatique et prosodiques, comme le suggèrent les modèles interactifs [Bol97], [Seg,00]. Dans la perspective plus large de la compréhension de la parole, il faudra s'attacher à montrer comment la prosodie participe à la fois au décodage proprement dit des énoncés, ce dernier donnant accès à la représentation de sa forme logique et à l'activation des processus inférentiels qui concourent à l'identification de son contenu propositionnel [Reb98]. Le modèle de la pertinence de [Spe89] offre un cadre propice à ce type de recherche. Dans cette perspective, nous poserons comme hypothèse de travail que la prosodie a un rôle central dans le processus cognitif d'intégration des différentes sources de connaissance linguistique [Bro00] qui préside à la compréhension de la parole.

La nécessité de plus en plus pressante d'étudier la prosodie dans des situations de communication authentiques, conduit inévitablement les chercheurs à s'intéresser aux fonctions interactionnelles de la prosodie. Ces dernières, qui sont complémentaires, dans formation des messages interindividuels, des fonctions structurales dont il vient d'être question (ou qui peuvent même entrer en conflit avec elles) et qui participent du jeu des négociations conversationnelles, s'inscrivent dans la quadruple perspective de « l'hétérogénéité de la parole » [Ker91], [Ber99], des théories de l'énonciation [Ker80], du modèle de l'interprète, ou de la « parole adressée »[Reb98] et du modèle de la contextualisation, ou du « contexte revisité » [Aue92], [Gum92] Bien que très prometteuses, notamment dans l'optique d'une application à la synthèse vocale du dialogue, les recherches sur les fonctions interactionnelles de la prosodie n'ont guère dépassé le stade d'une « approche en mosaïque ». Elles concernent principalement la poursuite des travaux sur la gestion des tours de parole [Cut86], l'étude de certains phénomènes temporaires - mais néanmoins prégnants - comme les régulateurs d'écoute, les marqueurs de co-énonciation [Jean99], d'inter-synchronie (par exemple, de « *pitch concord* »), de polyphonie [Gün99], ainsi qu'une réinterprétation de la théorie des actes illocutoires dans une visée plus dynamique que celle du modèle standard [Ghi93], [Pur98], (pour une étude plus générale touchant à différents aspects de la prosodie interactionnelle, on se reportera à [Gros91]).

L'expression de l'affect (le terme étant pris ici dans son acception étymologique) que l'on a longtemps considéré comme la fonction de base de la prosodie, recouvre un ensemble de phénomènes étroitement imbriquées, mais qu'il est indispensable de dissocier pour les besoins de l'analyse. Les premiers concernent des aspects éthologiques de la communication selon lesquels des distinctions prosodiques rudimentaires, comme mélodie montante vs. mélodie descendante ou registre haut vs registre bas, sont censées refléter des états psycho-physiologiques primaires, tels que l'antagonisme tension/détente, soumission, agressivité, etc. [Oha96] (ce qui a fait dire au linguiste Bolinger que la prosodie « *est*

incrustée dans une matrice de réactions instinctives »). Il est tentant d'interpréter ces distinctions comme les expressions vocales de la « métaphore haut/bas » qui constitue probablement l'un des substrats les plus vivaces du patrimoine cognitique humain [Lak80],

Pour de nombreux chercheurs, l'expression de l'affect fait uniquement référence à la manifestation des émotions. Ce domaine de prédilection des philosophes et des psychologues, suscite depuis peu un grand intérêt chez les phonéticiens et les linguistes. L'analyse prosodique se heurte cependant ici à trois sources de difficultés majeures : a) l'élaboration préalable d'une classification des émotions qui permette de distinguer, sur la base de critères explicites, entre émotions primaires (joie, colère, etc.), sentiments (tristesse) et émotions « socialisées » ou attitudes (doute, surprise) [Arn91], ces dernières s'apparentant à la fois aux fonctions modales et illocutoires avec lesquelles elles interférent (ce qui nous renvoie à la problématique des distinctions catégorielles et graduelles). Il est clair, toutefois, qu'une telle classification fait encore défaut [Léo93] ; b) la mise au point de protocoles expérimentaux adéquats ; et c) l'inventaire des marqueurs prosodiques qui participent à l'expression de l'affect [Sch91], [Moz98].

Enfin, l'expression de l'affect concerne également les phénomènes d'emphase qui procèdent à la fois de la manifestation des émotions et d'options conventionnelles par lesquelles le locuteur choisit de s'investir dans ses actes énonciatifs (fonction expressive et fonction de contextualisation) ou d'influencer son auditoire (fonction impressive). Envisagée sous ces angles divers, il devient évident que l'emphase constitue une des composantes majeures des échanges conversationnels dont l'analyse prosodique doit rendre compte [Sel94], [Bag00]

Les diverses fonctions que nous venons de passer en revue montrent à l'évidence que les signaux prosodiques sont hautement polysémiques, ce qui représente un obstacle majeur à l'explication des relations entre formes et fonctions qui doit être l'aboutissement d'une véritable « cognitique » de la prosodie. Pour parachever le tableau, il est indispensable d'évoquer, si l'on se place du point de vue de l'auditeur, une fonction prosodique qui ne participe pas à la compréhension proprement dite des messages, mais qui contribue cependant à l'interprétation des événements communicatifs dont ils sont porteurs. Il s'agit de la fonction indicielle selon laquelle les signaux prosodiques fonctionnent comme des « marqueurs attributifs », fournissant des indices à la fois sur les caractéristiques physiques du locuteur (telles que le sexe, la taille, le poids, l'état de santé), son état psychologique (niveau de « *stress* », par exemple) et son appartenance à une communauté dialectale [Coq00], [Mor96] et socioculturelle [Léo93]. Les travaux sur ce thème sont peu nombreux et il est certain que la prise en considération de la prosodie devrait apporter une contribution conséquente à la sociolinguistique.

La fonction indicielle, ou symptomatique de la prosodie concerne également la caractérisation des différents styles discursifs (cf. [Esk93] et[Ast99], pour des discussions récentes sur ce sujet), ainsi que les dysfonctionnements d'origine pathologique [Bha94], [Bau00]. Sur ce dernier point, l'élaboration d'une typologie informée des dysprosodies pourra constituer un auxiliaire précieux en vue de l'aide au diagnostic et de la mise en œuvre de méthodes de réhabilitation fondées sur la prosodie.

3. PROPOSITIONS

Dans la dernière partie de cet exposé, qui tiendra lieu de conclusion (ou plutôt de « non-conclusion »), nous ferons état de deux propositions concernant, d'une part, l'élaboration d'un cadre de référence adapté à la complexité fonctionnelle des objets prosodiques et, d'autre part, une définition explicite des niveaux d'analyse et de représentation qui concourent à établir des liens rigides entre les observables et les formes sonores qui sous-tendent les systèmes prosodiques.

3.1 Le cadre interprétatif

Afin d'ébaucher une approche intégrative qui permettrait de rendre compte à la fois de l'organisation et des diverses motivations des signes prosodiques, nous proposons le cadre interprétatif d'une « Grammaire Ecologique » [Dic00], le terme de grammaire étant pris dans la signification large de description des modes d'existence et de fonctionnement d'une langue naturelle ou, plus spécifiquement, de toute sémiotique langagière. L'association inhabituelle de la notion d'écologie à celle de grammaire signifie simplement que, pour l'étude de la prosodie, nous appréhendons le langage comme un éco-système, c'est-à-dire comme un mode d'expression qui s'adapte en permanence au milieu dans lequel il se déploie, en fonction de l'environnement cognitif changeant des usagers, des intentions des locuteurs et des pressions exercées par les forces interactionnelles qui régulent les échanges conversationnels. Telle que nous la concevons, la Grammaire Ecologique (ci-après GE).s'inscrit donc à la fois dans le paradigme d'une pragmatique de la communication (incluant des aspects qui relèvent de l'énonciation, de l'illocutoire, de la contextualisation et de l'expression de l'affect) et dans celui des sciences cognitives dont l'un des présupposés est que les capacités intellectuelles et les fonctions perceptives sont des systèmes de traitement de l'information en provenance de l'environnement [Cos98] . Pour être plus précis, la GE se présente comme une supra-grammaire constituée de deux sous-grammaires interfacées : a) une grammaire des « expressions linguistiques » (qui s'apparente à la grammaire selon Chomsy, mais qui comprend en fait la description et la représentation des primitives et des constructions des systèmes syntaxique, sémantique et phonologiques, ainsi que la spécification d'un ensemble de paramètres phonétiques) et b) une grammaire de la « contextualisation », ce terme étant pris dans la

signification large que lui attribuent Gumperz et ceux qui se réclament de sa conception [Aue92]. La grammaire de la contextualisation doit rendre compte d'une pragmatique de la pertinence, de la subjectivité et de l'inter-subjectivité dans le langage, c'est-à-dire, d'un ensemble de stratégies qui soient révélatrices de la « compétence pragmatique » (l'expression étant de Chomsy, lui-même !) des locuteurs, ces stratégies étant notamment celles par lesquelles les locuteurs/auditeurs :

- Manifestent ostensiblement leurs intentions communicatives et interprètent inférentiellement celles de leurs interlocuteurs ;
- Produisent des messages « communicativement bien formés », c'est-à-dire adaptés à leurs objectifs et à leurs contextes conversationnels ;
- Se mettent en scène personnellement dans les actes de langage qu'ils expriment (contextualisation de l'affect et des attitudes) ;
- Participent interactivement à la mise en scène et à la régulation des échanges conversationnels (ces stratégies étant soumises à la fois à des rites socioculturels et à des enjeux d'influence).

La GE se présente en définitive comme une démarche unificatrice qui s'efforce d'unir les approches le plus souvent mutuellement exclusives du formalisme et du fonctionnalisme [New98]. Ce faisant, elle convie à dialoguer un panel de modèles et de théories qui émanent de champs disciplinaires divers, mais qui présentent à nos yeux une évidente parenté conceptuelle ainsi qu'une féconde complémentarité. Nous citerons, à titre d'exemple, le modèle de la pertinence de [Spe89], le modèle psycholinguistique de la parole de [Lev89], la théorie de la grammaire fonctionnelle de [Dik89], la théorie de la structure informationnelle de [Lam94], le modèle H & H de Lindblom [Lin90], la théorie de l'optimalité de [Pri93], le modèle connexionniste de compétition de [Bat89], la théorie des actes de sens de [Bru90], le modèle de la logique des actes illocutoires de [Sea85], ainsi que les diverses approches de l'interaction verbale [Vio92], [Cos98].

Le cœur de la GE se situe à l'intersection des deux grammaires, car cette interface est un locus où s'articulent une linguistique « des états » et une linguistique « des opérations », ou mieux : où se confrontent les deux aspects essentiels des activités cognitives mises en jeu dans le langage : l'aspect représentationnel et l'aspect procédural mis au service de la dynamique communicative. L'interface doit donc rendre compte, dans la perspective d'une extension de la théorie de l'optimalité (qui dépasserait le cadre restrictif actuel limité à un jeu de contraintes phonologiques), de la façon dont les contraintes pragmatiques de la contextualisation et les contraintes linguistiques (notamment les contraintes syntaxiques et métriques) conspirent ou entrent en conflit dans le choix des configurations prosodiques signifiantes qui activent en ligne l'interprétation sémantique des messages.

Nous ne sommes pas en mesure d'approfondir ici, faute de place, une étude de cas pour illustrer une application de notre approche. Nous nous permettons cependant de renvoyer le lecteur à notre étude de l'accentuation du français [Dic99], dans laquelle nous mettons en pratique ces concepts théoriques et où nous montrons que les principes de « bipolarisation » et de « dominance accentuelle » peuvent rendre compte des dispositifs d'alignement qui régulent l'accentuation probabilitaire du français [Fón80]. Nous évoquerons également la problématique cruciale de la focalisation. Nous concevons la focalisation comme une opération pragmatique de sélection des entités sémantiques qui participent à la construction de l'échelle focale des messages. Cette construction repose à la fois sur une mise en perspective informationnelle des relations de forme à fond qui s'établissent dans le cadre d'un contexte pro-actif, et sur l'engagement épisodique des locuteurs dans l'expression de leurs messages. La mise en œuvre de la focalisation est également fondée sur une « stratégie de l'interprète » qui procède d'un calcul permanent de l'état cognitif de l'allocutaire et d'une adaptation à cet état. La construction de l'échelle focale utilise, outre des marqueurs lexicaux et syntaxiques, des signaux mimo-gestuels [Gai95] et des dispositifs prosodiques complexes (accentuation nucléaire mobile, bipolarisation accentuelle, marqueurs d'emphase, extension de la dynamique tonale, etc.) qui attestent de la plasticité de la langue en action.

3.2 Les niveaux d'analyse et de représentation.

Nous soutenons avec conviction qu'une théorie prosodique est tenue de spécifier les niveaux d'analyse et de représentation sur lesquels elle se fonde. La définition de ces niveaux doit être suffisamment explicite, afin qu'elle puisse éventuellement donner lieu à une « entreprise de falsification ».

Dans cette perspective, nous présentons [Hir98], [Hir00] un modèle de représentation à quatre niveaux dont les traits majeurs sont la réversibilité, la mise en œuvre d'un principe d'interprétabilité (qui stipule que tout niveau de représentation doit être interprétable aux niveaux adjacents) et la caractère automatique des processus de transcodage.

En ce qui concerne plus précisément les représentations substantielles et formelles de l'intonation, nous distinguons ainsi (figure 1), du niveau le plus concret au niveau le plus abstrait :

- *le niveau physique* (a). C'est le niveau où la manifestation concrète de l'intonation est représentée de façon analytique par les configurations « brutes » de Fréquence fondamentale (F0).
- *Le niveau phonétique* (b). Abstraction faite des contraintes universelles de production et de perception, les configurations brutes du niveau

physique sont interprétables comme des représentations phonétiques de l'intonation. Ces représentations phonétiques (obtenues au moyen de l'algorithme MoMel, [Hir00]) ont la forme de courbes lisses et continues, constituées de séquences de points-cibles reliés par une fonction d'interpolation monotone.

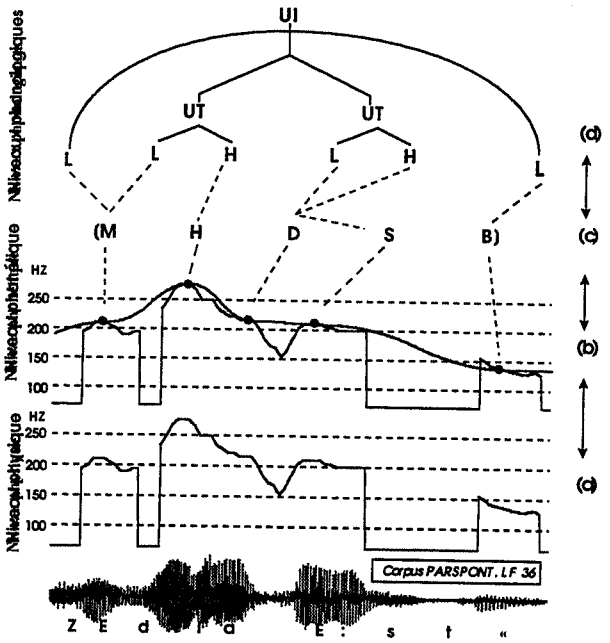


Figure 1 : Illustration des niveaux d'analyse et de représentation de l'intonation de l'énoncé « J'ai dit la veste ».

- *Le niveau phonologique de surface* (c). Il donne lieu à une représentation en termes de séquences de segments tonals correspondant à un codage discret des cibles de la représentation phonétique. Ce codage est réalisé au moyen du système de transcription INTSINT [Hir98], qui est constitué des huit symboles suivants : M(id), T(op) et B(ottom), pour la notation des registres ; H(igher), L(ower), S(ame) ; D(ownstepep) ; U(pstep), pour la notation des valeurs mélodiques relatives à l'intérieur des registres.

- *Le niveau phonologique sous-jacent* (d). Ce niveau de représentation, qui est dépendant de la théorie, spécifie les primitives et les constructions qui forment le système intonatif noyau de la langue étudiée. Pour ce qui est du français, ce système noyau est représenté sous la forme de schèmes tonals associés aux deux constituants prosodiques de base que sont l'Unité Tonale (UT), avec son schème LH et l'Unité Intonative (UI) , avec ses schèmes LL et LH.

Dans la démarche ascendante, le passage de la représentation analytique à la représentation phonologique de surface peut s'effectuer selon une procédure d'étiquetage manuelle ou automatique. Dans la démarche descendante, la représentation phonologique de surface est dérivable de la représentation sous-jacente par un ensemble de règles qui ont été décrites dans [Hir86] et [Dic96]. Il est également possible, dans le cas d'une application à la synthèse vocale, de générer une courbe de fréquence fondamentale à partir de la représentation phonologique de surface.

BIBLIOGRAPHIE

[Arn91] Arndt, H. & Janney, W. (1991). Verbal, prosodic and kinesic emotive contrasts in speech. Journal of Pragmatics, 15, 521-549.

[Ast99] Astesano C. (1999) Rythme et discours : Invariance et sources de variabilité des phénomènes accentuels en français, Thèse de Doctorat. Université de Provence.

[Aub91] Aubergé, V. (1991). La synthèse de la parole : des règles au lexique, Thèse de Doctorat, Université Stendhal, Grenoble.

[Aue92] Auer, P. & Di Luzio, A. (1992). The contextualization of Language. John Benjamins.

[Aue93] Auer, P. (1993). Is a rhythm-based typology possible ? A study of the role of prosody in phonological typology. Universität Hamburg.

[Bac86] Bacri, N. (1986). Fonctions de l'intonation dans l'organisation perceptive de la parole. Thèse d'Etat, Paris VIII.

[Bag00] Bagou, O. & Di Cristo, A. (2000). L'implication emphatique dans la narration orale spontanée : validation perceptive et réalisations acoustiques. XIII^{èmes} Journées d'Etudes sur la Parole. Ce volume.

[Bal99] Balan, A. & Gandour, J.(1999). Effect of sentence length on the production of linguistic stress by left-and right-hemisphere-damaged patients, Brain and Language, 67, 73-94.

[Bar94] Barbosa, P. & Bailly, G. (1994). Characterization of rhythmic patterns for text-to-speech synthesis, Speech Communication, 15, 127-137.

[Bat89] Bates, E. & MacWhinney, B. (1989). Functionalism and the competition model. In MacWhinney, B. & Bates, E. (eds.). The Crosslinguistic Study of Sentence Processing. Cambridge University Press, 3-73.

[Bau97] Baum, S.R., Pell, M.D., Leonard, C.L. & Gordon, J.K. (1997). The ability of right-and Left-hemisphere-damaged individuals to produce and interpret prosodic cues marking phrasal boundaries. Language and Speech, 40, 313-330.

[Bau00] Baum, S.R. & Pell, M.D. (2000). The neural bases of prosody : insights from lesion studies ans neuroimaging. Aphasiology (forthcoming).

[Bec86] Beckman, M. & Pierrehumbert, J. (1986). Intonational Structure in Japanese and English, Phonology Yearbook, 3, 255-309.

[Bec96] Beckman, M. (1996). The parsing of prosody. Language and Cognitive Processes, 11, 17-67.

[Ber89] Bertinetto, P.M. (1989). Reflections on the dichotomy 'stress' vs 'syllable-timing'. Revue de Phonétique Appliquée, 91/93, 99-130.

[Ber90] Berrindonner, A. (1990). Pour une macro-syntaxe. Travaux de Linguistique (gand), 21, 25-36.

[Ber93] Berrendonner, A. (1993). Périodes. In Parret, H. (ed.). Temps et Discours, Presses Universitaires de Louvain, 47-61

[Ber99] Bertrand, R. (1998). De l'hétérogénéité de la parole : analyse énonciative de phénomènes prosodiques et kinésiques dans l'interaction interindividuelle. Thèse de Doctorat. Université de Provence.

[Bha94] Bhatt, P. (1994). Pathologies des systèmes intonatifs. Calap, 11, 63-85.

[Bla95] Blaauw, E. (1995). On the perceptual classification of spontaneous and read speech. OTS, Utrecht University.

[Bla90] Blanche-Benveniste, C. (1990). Le français parlé : études grammaticales. Ed. Du CNRS.

[Bol51] Bolinger, D.L. (1951). Intonation : levels vs configurations. Word, 7, 199- 210.

[Bol58] Bolinger, D.L. (1958). A theory of pitch accent in english. Word, 14, 109-149.

[Bol97] Boland, J.E. & Cutler, A. (1996). Interaction with autonomy : multiple outputs models and the inadequacy of the Great Divide. Cognition, 58, 309-320.

[Bou99] Boudouresques, N. (1999). Evaluation et analyse pluriparamétrique des troubles prosodiques observés chez des patients traumatisés crâniens. Mémoire de DEA, Université de Provence.

[Bou99] Boudon, P. (1999). Le réseau du sens. Peter Lang.

[Bro00] Brown, C.M., Hagoort, P. & Kutas, M. (2000). Postlexical integration processes in language comprehension : evidence from brain-imaging research. In Cazzaniga, M. (ed.).

[Bru77] Bruce, G. (1977). Swedish word accents in sentence perspective. Gleerup. Lund.

[Bru90] Bruner, J. (1990). Acts of Meaning. Harvard University Press.

[Cae91] Caelen-Haumont, G. (1991). Analyse des interactions entre modèles syntaxiques, sémantiques, pragmatiques et paramètres prosodiques. Thèse de Doctorat d'Etat.

[Cah97] Cahiers de Linguistique Française (1997). Problèmes d'analyse du discours. Vol. 19. Université de Genève.

[Caz00] Cazzaniga, M.S. (2000). The New Cognitive Neurosciences. MIT Press.

[Cho68] Chomsky, N. & Halle, M. (1968). The Sound Pattern of English. Harper & Row, New York.

[Cho81] Chomsky, N. (1981). Lectures on Government and Binding. Foris, Dordrecht.

[Col99] Colas, A. (1999). Introducing infants to referential events : a development study of maternal ostensive marking in French. Journal of Child Language, 26, 113-131.

[Coq00] Coquillon, A. (2000). Marseillais et toulousains gèrent-ils différemment leurs pieds ? Caractéristiques prosodiques du schwa dans les parlers méridionaux. XIII^{èmes} Journées d'Etudes sur la Parole. Ce volume.

[Cos98] Cosnier, J. (1998). Le retour de psyché. Desclée de Brouwer.

[Chr94] Christophe, A. Dupoux, E., Bertoncini, J. & Mehler, J. (1994). Do infants perceive word boubdaries ? An empirical study of the bootstrapping of lexical acquisition. Journal of the Acoustical Society of America, 95, 1570-1580.

[Cut86] Cutler, A. & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In John-Lewis, C. (ed.). Intonation in discourse. Croom Helm, 139-155.

[Cut97] Cutler, A., Dahan, D. & Van Donselaar, W. (1997). Prosody in the comprehension of the spoken language. Language and Speech, 40, 141-201.

[Cut99] Cutler, A.(1999). Prosodic structure and word recognition. In Friederici, A. (ed.). Language Comprehension. Springer, 41-70.

[Dar75] Darwin, C. (1975). On the dynamic use of prosody in speech perception. Haskins Labs Stat. Rep. On Speech Research, 42/43, 103-115.

[Dau83] Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51-62.

[Deb96] De Boysson-Bardies, B. (1996]. Comment la parole vient aux enfants. Odile Jacob, Paris.

[Del95] Delais-Roussarie, E. (1995). Pour une approche parallèle de la structure prosodique. Etude de l'organisation prosodique et rythmique de la phrase française. Thèse de Doctorat. Université de Toulouse-Le-Mirail.

[Del00] Delais-Roussarie, E. (2000). Vers une nouvelle approche de la structure prosodique. Langue Française (à paraître).

[Del66] Delattre, P. (1966]. Les dix intonations de base du français. The French Review, 40 (1), 1-14.

[Dic85] Di Cristo, A. (1985]. De la microprosodie à l'intonosyntaxe. Publications de l'Université de Provence.

[Dic86] Di Cristo, A. & Hirst, D.J. (1986). Modelling French micromelody : analysis and synthesis. Phonetica, 43-11-30.

[Dic93] Di Cristo, A. & Hirst, D.J. (1993). Prosodic regularities in the surface structure of French questions. Working Papers (Lund University), 41, 268-271.

[Dic96] Di Cristo, A. & Hirst, D.J. (1996). Vers une typologie des unités intonatives du français. Actes des XXIes JEP, 219-222.

[Dic99] Di Cristo, A. (1999). Le cadre accentuel du français : essai de modélisation. Langues, 2 (3), 184-205 et Langues, 2 (4), 258-269.

[Dic97] Di Cristo, A., Di Cristo, P. & Véronis, J. (1997). A metrical model of rhythm and intonation for French text-to-speech synthesis. In Botinis, A. (ed.). Intonation : Theory, Models and Applications (ESCA), 83-86.

[Dic00] Di Cristo, A. (2000). Une grammaire écologique comme cadre interprétatif de la prosodie de la parole.Communication au Congrès International de Sémiotique d'Imatra (Finl.), juin 2000.

[Dik89] Dik, S. (1989). The Theory of Functional Grammar. Foris.

[Dra98] Psychological processes involved in the temporal oganization of complex auditory sequences, Music Perception, 16, 11-26.

[Due87] Duez, D. (1987). Contribution à l'étude de la structuration temporelle de la parole en français. Thèse de Doctorat. Universite de Provence.

[Dur90] Durand, J. (1990). Generative and Non-Linear Phonology, Longman, London.

[Esk93] Eskénazi, M.(1993). Trends in speaking styles research. Proceedings Eurospeech 93 (Berlin), 501-505.

[Ess88] Esser, J. (1988). Comparing reading and speaking intonation. Rodopi. Amsterdam.

[Fer89] Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants : is the melody the message ? Child Development, 60, 1497-1510.

[Fod83] Fodor, J.A. (1983). The Modularity of Mind. MIT Press.

[Fón79] Fónagy, I. (1979). La fonction prédictive de l'intonation. In Léon, P. & Rossi, M. (eds.). Problèmes de prosodie. Didier, 112-120.

[Fón80] Fónagy, I. (1980). L'accent en français : accent probabilitaire. In Fónagy, I. & Léon, P. L'Accent en français contemporain, Studia Phonetica, 13, 123-233.

[Fou99] Fougeron, C. (1999). Prosodically conditioned articulatory variations :a review. UCLA working Papers in Phonetics, 97, 1-73.

[Fuc92] Les linguistiques contemporaines. Hachette, Paris.

[Fuj79] Fujisaki, H., Hirose, K. & Ohta, K. (1979). Acoustic features of the fundamental frequency contours of declarative sentences in Japanese. Ann. Bull. Res. Inst. Logopedics and Phoniatrics, 13, 163-173.

[Gar65] Garde, P. (1965). Accentuation et morphologie. La Linguistique, 2, 25-39.

[Ghi93)] higlione, R. & Trognon, A. (1993). Où va la pragmatique. Presses Univ. de Grenoble.

[Gua91] Guaïtella, I. (1991). Etude des relations entre geste et prosodie à travers leurs fonctions rythmique et symbolique. XIIth ICPhS, 266-269.

[Gua95] Guaïtella, I. (1995). Mélodie du geste, mimique vocale. Semiotica, 103, 253-276.

[Gol76] Goldsmith,J. (1976). Autosegmental Phonology. PhD. Thesis. MIT.

[Gol90] Goldsmith, J.(1990). Autosegmental and Metrical Phonology. Blackwell.

[Gra98] Grabe, E., Nolan, F. & Farrar, K. (1998). A comparative transcription system for intonational variation in English, Proceedings ICLSP 5, Sydney.

[Grø95] Grønnum, N. (1995). Superposition and subordination in intonation : a non-linear approach. Proc. XIIIth ICPhS, 2, 124-131.

[Gro91] Grosjean, M. (1991). Les musiques de l'interaction. Thèse de Doctorat. Un. de Lyon II.

[Gro92] Grosz, B. & Hirchberg, J. (1992). Some intonational characteristics of discourse structure. Proc. ICSLP (Banff), 429-432.

[Gül95] Gülich, E. & Kotschi, T. (1995). Discourse production in oral communication. In Quasthoff, M. (ed.). Aspects of Oral Communication. De Gruyter, 30-66.

[Gum92] Gumperz, J.(1992). Contextualization and understanding. In Duranti, A. & Goodwin, C. (eds.). Rethinking Context. Language as an Interactive Phenomenon. Cambridge University Press, 229-252.

[Gün99] Günthner, S. (1999). Polyphony and the 'layering of voices' in reported dialogues. Journal of Pragmatics, 31, 685-708.

[Gus99] Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts. Language and Speech, 42,, 283-305.

[Hal87] Halle, M. & Vergnaud, J.R. (1987). An Essay on Stress. MIT Press.

[Hal95] Halle, M. & Isdardi, W. (1995). Stress and Metrical Structure. In Goldsmith, J. (ed.). The Handbook of Phonological Theory. Blackwell, 403-443.

[Har90] 't Hart, J., Collier, R. & Cohen, A. (1990). APerceptual Study of Intonation. Cambridge University Press.

[Hau94] Hauser, M. & Anderson, K. (1994). Left hemisphere dominance for processing vocalizations in adult but not infant, rhesus monkeys ? proceedings of the National Academy of Sciences, 91, 3946-3948.

[Hay81] Hayes, B. (1981). A Metrical Theory of Stress Rules. PhD. Thesis. MIT.

[Hay90] Hayes, B. (1995). Metrical Stress Theory. The University of Chicago Press. Chicago.

[Haz83] Hazael-Massieux, M.C. (1983). Le rôle de l'intonation dans la définition et la structuration de l'unité de discours. BSL, 78, 99-160.

[Hir83] Hirst, D.J. (1983). Structures and categories in prosodic representations. In Cutler, A. & Ladd,

R. (eds.). Prosody : Moels and Measurements. Springer, 93-109.

[Hir84] Hirst, D.J. & Di Cristo, A. (1984). French intonation : a parametric approach. Die Neuren Sprachen, 83 (5), 554-569.

[Hir87] Hirst, D.J. (1987). La description linguistique des systèmes prosodiques : une approche cognitive. Thèse de Doctorat d'Etat. Université de Provence.

[Hir86] Hirst, D.J. & Di Cristo, A. (1986). Unités tonales et unités rythmiques dans la représentation de l'intonation. XV èmes JEP (Aix-enProvence), 93-95.

[Hir98] Hirst D.J. & Di Cristo A (19 98) Intonation systems : a survey of twenty languages, Cambridge University Press.

[Hir00] Hirst, D.J., Di Cristo, A. & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (ed.). Prosody : Theory and Experiments. Kluwer Academic Press.

[Ind00] Indefrey, P. & Levelt, W.J.M. (2000). The neural correlates of language production. In Gazzaniga, M.S. (ed.). The New Cognitives Neurosciences, MIT Press, 845-865.

[Ise98] Isel, F. & Bacri, N. (1998). Segmentation en mots et compétitions lexicales. XXIIèmes Journées d'Etudes sur la Parole, 41-48.

[Jea99] Jeanneret, M.T. (1999). La coénonciation en français. Peter Lang.

[Jes98] Jescheniak, J.D., Hahne, A. & Friederici, A. (1998). Brain activity patterns suggest prosodic influences on syntactic parsing in the comprehension of spoken sentences, Music and Perception, 16, 55-62.

[Jus97] Jusczyk, P. (1997). The Discovery of Spoken Language. MIT Press.

[Ker80] Kerbrat-Orecchioni, C. (1980). L'Enonciation : de la subjectivité dans le langage. A. Colin.

[Ker91] Kerbrat-Orecchioni, C. (1991). Hétérogénéité énonciative et conversation. In Parret, H. (ed.). Le sens et ses hétérogénéités. Ed. Du CNRS, 121-138.

[Kin94] Kingston, J. & Diehl, R.L. (1994). Phonetic knowledge, Language 70, 419-454.

[Kon97] Konopczynski, G. (1997). Developmental interactive intonology : theory and applications. A paraître dans Lynch, M. (ed.). The Cognitive Science of Prosody. North Holland, Amsterdam.

[Lac99] Lacheret-Dujour A. & Beaugendre, F. (1999). La prosodie du français. Ed. Du CNRS.

[Lad86] Ladd, R. (1986). Intonational phrasing : the case for recursive prosodic structure. Phonology Yearbook, 3, 311-340.

[Lad90] Ladd, R. (1990). Metrical representation of pitch register. In Kingston, J. & Beckman, M. (eds.). Papers in Laboratory Phonology 1, 35-57.

[Lad93] Ladd, R. (1993). Notes on the phonology of prominence, Working Papers 41 (Lund), 10-15.

[Lad95] Ladd, R. (1995). « Linear » and « overlay » descriptions : an autosegmental-metrical middle way. Proc.XIIIth ICPhS (Stokolm), 2, 116-123.

[Lad96] Ladd, R. (1996). Intonational Phonology. Cambridge University Press. Proc. XIIIth ICPhS, 2, 116-123.

[Lak80] Lakoff, G. & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.

[Lak93] Laks, B. (1993). La constituance revisitée. In Laks, B. & Plénat, M. (eds.) De natura sonorum. Presses Universitaires de Vincennes, 173-220.

[Lam94] Lambrecht, K. (1994). Information structure and sentence form. Cambridge University Press.

[Lav91] Laver, J. (1991). The Gift of Speech. Edinburgh University Press.

[Leb73] Leben, W. (1973) Suprasegmental Phonology. PhD Thesis. MIT.

[Leh70] (Lehiste, I. (1970). Suprasegmentals. MIT Press.

[Léo93] Léon, P. (1993). Précis de phonostylistique. Nathan.

[Lev89] Levelt, W.J.M. (1989). Speaking. MIT Press.

[Lib75] Liberman, M. (1975). The Intonational System of English. PhD. Thesis. MIT.

[Lib92] Liberman, M. & Church, K. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S. & Sondhi, M. (eds.), Advances in Speech Signal Processing, Dekker, 791-831.

[Lin90] Lindblom, B. (1990). Explaining phonetic variation : a sketch of the H & H theory. In Hardcastle, W. & Marchal, A. (eds.). Speech Production and Speech Modelling. Kluwer, 403-440.

[Lin99] Lindfield, K.C., Wingfield, A. & Goodglass, H. (1999). The role of prosody in the mental lexicon. Brain and Language, 68, 312-317.

[Lou00] Louis, M., Di Cristo, A., Habib, M. & Hirst, D. (2000). Etude phonétique (segmentale et prosodique) d'un cas de jargon phonémique. XIII^{èmes} Journées d'Etudes sur la Parole. Ce volume.

[Luc90] Luce, P., Pisoni, D.& Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In Altmann, G.T. (ed.). Cognitive Models of Speech Processing, MIT Press, 122-147.

[Mar87] Marek, B. (1987). The Pragmatics of Intonation, Redakcja Wydawnictw KUL, Lublin.

[Mar72] Martin, J. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior, Psychological Review, 79, 487-509.

[Mar87] Martin, P. (1987). Prosodic and rhythmic structures in French. Linguistics, 25, 925-949.

[Mar90] Marslen-Wilson, W.D. (1990). Activation, competition and frequency in lexical acces. In Altmann, G.T. (ed.). Cognitive Models of Speech Processing, MIT Press, 148-172.

[Mar92] Marslen-Wilson, W.D., Tyler, L.K., Warren, P., Grenier, P. & Lee, C.S. (1992). Prosodic effects in minimal attachment, The Quarterly Journal of Experimental Psychology, 45, 73-87.

[Meh00] Mehler, J. & Christophe, A.(2000). Acquisition of Languages : infant and adult data. In Gazzaniga, M.S. (ed.). The New Cognitive Neurosciences, MIT Press, 897-908.

[Mer87] Mertens, P. (1987). L'intonation du français. De la description linguistique à la reconnaissance automatique. Thèse de Doctorat. KU Leuven.

[Mne92] McNeil, D. (1992). Hand and Mind : What Gestures Reveal about Thought. University of Chicago Press.

[Möb95] Möbius, B. (1995]. Components of a quantitative model of German intonation. Proc. XIIIth ICPhS, 2, 108-115.

[Mon93] Monnin, P. & Grosjean, F. (1993). Les structures de performance en français : caractérisation et prédiction. L'Année Psychologique, 93, 9-30.

[Mor98] Morel, M.A. .& Danon-Boileau, L. (1998). Grammaire de l'intonation, l'exemple du français. Ophrys. Paris.

[Mor96] Mora, E. (1996). Caractérisation prosodique de la variation dialectale de l'espagnol parlé au Venezuela. Thèse de Doctorat. Un. de Provence.

[Mox93] M. & Sandford, A. (1993). Communicating quantities : a psychological perspective. Erlbaum.

[Moz98] Mozziconacci, S. (1998). Speech Variability and Emotion. Thèse de Doctorat. Eindhoven.

[Naz98] Nazzi, T., Bertoncini, J. & Mehler, J. (1998). Language discrimination by newborns : Towards an understanding of the role of rhythm. Journal of Experimental Psychology, 24, 1-11.

[Nes86] Nespor, M. & Vogel, I. (1986). Prosodic Phonology, Foris. Dordrecht.

[Nes96] Nespor, M., Guasti, T . & Cristophe, A. (1996). Selecting word order : the rhythmic activation principle. In Kleinhenz, U. (ed.) Interfaces in Pgonology. Akademie Verlag, Berlin, 1-26.

[New98]Newmeyer, F.J. (1998). Language Form and Language Function. MIT Press.

[Nic96] Nicol, J.L. (1996). What can prosody tell a parser ? Journal of Psycholinguistic Research, 25, 179-192.

[Nie98] Niemi, J. (1998). Modularity of prosody : autonomy of phonological quantity and intonation in aphasia, Brain and Language, 61, 45-63.

[Oha96] Ohala , J. (1996). Ethological theory and the expression of emotion in the voice. ICSLP 1996, 1812-1815.

[Pos00] Post, B. (2000). Tonal and Phrasal Structures in French Intonation. Thesus. The Hague.

[Pier80] Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation. PhD. Thesis. MIT.

[Pin93] Pinker, S. (1993). Language acquisition. In Posner, M.I. (ed.). Foundations of Cognitive Science. MIT Press, 359-398.

[Pri93] Prince, A. & Smolensky, P. (1993). Optimality theory : constraint interaction in generative grammar (TR-2). Rutgers University, New Brunswick.

[Pur98] Purson, A. & Di Cristo, A. (1998). Aspects pragmatiques et prosodiques de la demande de confirmation en français. TIPA, 18, 113-126.

[Pyn98] Pynte, J. (1998). The role of prosody in semantic interpretation, Music Perception, 16, 79-97.

[Ram99] Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73, 265- 292.

[Reb98] Reboul, A. & Moeschler, J. (1998). Pragmatique du discours. .A. Colin, Paris.

[Ros81] Ross, E. (1981). The aprosodias : functional/ anatomical organization of the affective components of langage in the right hemisphere. Archives of Neurology, 38, 561-569.

[Ros81] Rossi, M., Di Cristo, A., Hirst, D.J., Martin, P. & Nishinuma, Y. (1981). L'intonation : de l'acoustique à la sémantique. Klincksieck.

[Ros99] Rossi, M. (1999). L'intonation, le système du français. Ophrys. Paris.

[Ros98] Rossi, M. & Peter-Defare, E. (1998). Les lapsus. Presses Universitaires de France.

[Sab96] Sabio, F.(1996). Description prosodique et syntaxique du discours en français : données et hypothèses. Thèse de Doctorat.

[San99] Sandler, W. (1999). Prosody in two natural language modalities. Language and Speech, 42, 127-142.

[Sch91] Scherer, K.R. (1991). Emotion expression in speech and music. In Sundberg, L. et al. (eds.). Music, Language, Speech and Brain. MacMillan, 146-156.

[Sci95] Sciences Humaine (1995). Le langage sert-il à communiquer ? N° 51, Juin 1995.

[Sea85] Searle, J. & Vanderveken, D. (1985). Foundations of Illocutionary Logic. Cambridge University Press.

[Seg00] Segui, J. & Ferrand, L. (2000). Leçons de parole. Odile Jacob.

[Sel78] Selkirl, E.O. (1978). On prosodic structure and its relation to sybtactic strcucture. In Fretheim, T. (ed.). Nordic Prosody II, Tapir.

[Sel84] Selkirk E.O.(1984). Phonology and Syntax : The Relation Between Sound and Structure. MIT Press.

[Sel86] Selkirk, E.O. (1986). On derived domains in sentence phonology. Phonology Yearbook, 3, 371-405.

[Sel95] Selkirk, E.O. (1995). The prosodic structure of function words. In Beckman, J., Urbanczyk, S. & Walsh, L. (eds.). Ptimality Theory Occasional Papers, UMOP 18, UMASS /Amherst, 439-470.

[Sel94] Selting, M. (1994). Emphatic speech style- with special focus on the signalling of heightened emotive involvement in conversation. Journal of Pragmatics, 22, 375-408.

[Sha96] Shattuck-Hufnagel, S. & Turk, A.E. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. Journal of Psycholinguistic Research, 25 (2), 193-247.

[Sil92] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, Wightman, C.M., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). ToBI : a standard for labeling English prosody. Proceedings ICSLP 2, 867-870.

[Spe89] Sperber, D. & Wilson, D. (1989). La pertinence. Editions de Minuit.

[Ste00] Steinhauer, K., Alter, K. & Friederici, A.D (2000). Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nature (à paraître).

[Swe93] Swerts, M. (1993). Prosodic Features of Discourse Units. Doctoral Thesis. Eindhoven.

[Swi98] Swinney, D. & Love, T. (1998). The processing of discontinuous dependencies in language and music, Music Perception, 16, 63-78.

[Tal93] Tallal, P., Miller, S. & Holly Fitch, R. (1993). Neurobiological basis of speech : a case for the preeminence of temporal processing, Annals of the NY academy of Sciences, 682, 27-47.

[Tru95] Truckenbrodt, H. (1995). Phonological Phrases : their Relation to Syntax, Focus, and Prominence, PhD Dissertation. MIT.

[Vai74] Vaissière, J. (1974). On French prosody. MIT Quarterly Progress Report 114, 212-223.

[Vai75] Vaissière, J. (1975). Further note on French prosody. MIT Quarterly Progress Report 115, 251-261.

[Vai83] Vaissière, J. (1983). Language-independent prosodic features. In Cutler, A. & Ladd, R. (eds.). Prosody : Models and measurements, Springer-Verlag, 53-66.

[Val92] Vallduví, E. (1992). The Informational Component. Garland.

[Van89] Vandepitte, S. (1989). A pragmatic function of intonation. Lingua, 79, 265-297.

[Vih96] Vihman, M. (1996). Phonological Development. Blackwell.

[Vio92] Vion, R. (1992). La communication verbale. Hachette.

[Wen82] Wenk, B. & Wiolland, F. (1982). Is French really syllable-timed ? Journal of Phonetics, 10, 193-216.

[Whe97] Wheeldon, L. & Lahiri, A. (1997). Prosodic units in speech production. Journal of Memory and Language, 37, 356-381.

[Win75] Wingfield, A. (1975). The intonation-syntax interaction ; prosodic features in perceptual processing of sentences. In Cohen, A. & Nooteboom, S. (eds.) Structure and Process in Speech Perception. Springer-Verlag, 146-160.

[Yul80] Yule, G. (1980). Speakers'topics and major paratones, Lingua, 52, 33-47.

Systèmes de reconnaissance à grands vocabulaires : Progrès et défis

Jean-Luc Gauvain

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay, France
gauvain@limsi.fr
http://www.limsi.fr/tlp

ABSTRACT

The last decade has witnessed substantial advances in speech recognition technology, which when combined with the increase in computational power and storage capacity, has resulted in a variety of products already or soon to be on the market. This paper is a review of the state-of-the-art in large vocabulary continuous speech recognition, with a view towards highlighting recent advances. It also highlights issues in moving towards applications, discussing system efficiency, portability across languages and tasks, enhancing the system output by adding tags and non-linguistic information. Current performance in speech recognition and outstanding challenges for various applications are discussed.

1. INTRODUCTION

Ces dix dernières années, la reconnaissance automatique de la parole continue à grands vocabulaires a été un des domaines de recherche centraux en RAP, servant de banc d'essai pour évaluer modèles et algorithmes. L'intérêt pour cette technologie va bien au delà des systèmes de dictée de textes. Elle peut par exemple être employée pour l'accès vocal à des bases de données, ou l'indexation par le contenu de documents audiovisuels. Les progrès dans ce domaine profitent également à d'autres technologies, telles que la reconnaissance du locuteur et de la langue, qui utilisent les mêmes modèles. La reconnaissance de la parole concerne principalement la transcription d'un signal vocal en une suite de mots. La plupart des systèmes repose sur une modélisation statistique du processus de génération de la parole. De ce point de vue, le message est produit par un modèle linguistique qui estime $\Pr(w)$ pour toutes les suites de mots w, et le canal acoustique, encodant le message w dans le signal x, est modélisé par une densité de probabilité $f(x|w)$. Le décodage de la parole consiste alors à maximiser la probabilité *a posteriori* de w, ce qui est équivalent à maximiser le produit $\Pr(w)f(x|w)$. Les principes de base sur lesquels la plupart des systèmes sont fondés sont connus depuis de nombreuses années, c.-à-d. l'application de la théorie de l'information à la reconnaissance de la parole [5, 48], la représentation spectrale du signal vocal [26, 27], la programmation dynamique pour le décodage [93, 94], et l'utilisation de modèles acoustiques en contexte [17, 57, 86]. Malgré cela des progrès considérables ont été faits ces dernières années, en particulier pour la modélisation acoustique et le décodage. Ces progrès sont liés à la disponibilité de grands corpus de parole et de textes ainsi qu'aux puissances de calcul accrues qui ont permis le développement de modèles et d'algorithmes toujours plus complexes.

Cet article présente les avancées récentes de l'état de l'art, et explore des domaines d'application rendus possibles par ces progrès technologiques. Une importante avancée est la capacité des systèmes actuels à traiter des données non homogènes, par opposition à des enregistrements soigneusement préparés. Ceci est illustré par le traitement de documents radio ou télédiffusés: avec de nombreux changements de locuteurs, de conditions acoustiques, de thèmes, voire de langues. De nombreux progrès y ont contribué : une analyse acoustique plus robuste, des techniques d'apprentissage tirant profit de très grands corpus audio et textuels, des algorithmes de segmentation du flux audio, l'adaptation non supervisée des modèles acoustiques, des décodeurs plus performants avec des modèles linguistiques d'ordre plus élevés, et la capacité de traiter des vocabulaires beaucoup plus grands que par le passé (65k mots ou plus). Le développement de systèmes dans le cadre d'applications réelles (hors laboratoire) implique de reconsidérer certaines solutions, telles que l'enregistrement du signal, la compensation du bruit et du canal de transmission, et la capacité de rejet, tout en tenant compte des contraintes matérielles [34]. Les techniques mises en avant dans cet article ont été choisies en fonction de résultats expérimentaux obtenus dans différents laboratoires sur des données publiquement disponibles avec des systèmes au niveau de l'état de l'art.

2. MODÉLISATION ACOUSTICO-PHONÉTIQUE

La plupart des systèmes utilisent des modèles de Markov cachés (MMC) pour la modélisation acoustique [6, 23, 28, 35, 47, 61, 62, 67, 75, 77, 81, 97, 100]. D'autres utilisent des modèles segmentaux [41, 70, 105] ou des réseaux neuronaux [1, 11, 45] pour l'estimation des vraisemblances acoustiques, cependant tous les systèmes se servent du cadre des MMC pour combiner l'information linguistique et acoustique dans un seul réseau représentant le langage de l'application. Pour les systèmes fondés sur des MMC, le modèle est une densité de probabilité sur une séquence de vecteurs acoustiques. Les paramètres des vecteurs acoustiques sont choisis afin de réduire la complexité du modèle tout en essayant de garder l'information appropriée, c.-à-d. l'information linguistique. La plupart des systèmes utilisent des cepstres à court terme obtenus par transformée de Fourier ou via un modèle de prédiction linéaire. Les deux jeux de paramètres les plus util-

isés sont des coefficients cepstraux obtenus avec une analyse de type MFCC [19] ou avec une analyse PLP [44]. Dans les deux cas un spectre de puissance à court terme (20 à 30 ms) est estimé sur une échelle MEL, avec une période le plus souvent égale à 10 ms. Les deux jeux de paramètres ont été utilisés avec succès, mais l'analyse PLP s'avère plus robuste en présence de bruit pour certains systèmes [53, 98].

Les modèles de phones en contexte (triphones ou pentaphones) sont aujourd'hui les unités acoustiques les plus répandues. Comparées à des unités plus grandes telles que les diphones, les demisyllabes ou les syllabes, les modèles de phones en contexte offrent un plus large spectre de dépendances contextuelles avec la possibilité d'un mécanisme de repli vers des contextes fréquents. Le choix de l'ensemble des contextes modélisés est habituellement le résultat d'un compromis entre résolution et robustesse, et dépend fortement des données d'apprentissage disponibles. Ce qui est vraiment essentiel c'est d'ajuster le nombre de paramètres du modèle à la quantité de données d'apprentissage. Une technique très efficace pour limiter le nombres de paramètres des modèles sans sacrifier la résolution, consiste à tirer profit de la similitude entre certains états des MMC d'un même phonème en liant les distributions de ces états. Cette idée fondamentale est utilisée dans la plupart des systèmes avec de légères différences dans la mise en œuvre et dans le nom donné à ces groupes d'états (*senones* [46], *genones* [22], *PELs* [9], *tied-states* [103]). Ce partage de paramètres permet bien entendu de réduire la taille du modèle. Il peut être appliqué à tous les niveaux [90, 99] du modèle (allophone, état MMC, et gaussienne) mais plus de flexibilité est disponible au niveau des gaussiennes, où de grandes réductions peuvent être obtenues sans sacrifier les performances.

3. Modélisation lexicale

Le dictionnaire de prononciations est le lien entre le modèle acoustique et les entrées lexicales du modèle de langage, chaque entrée lexicale étant décrite comme une suite d'unités phonémiques. La conception d'un tel dictionnaire nécessite d'une part la sélection des éléments du vocabulaire en minimisant le nombre de mots hors vocabulaire, et d'autre part la détermination des prononciations possibles de chaque mot de ce vocabulaire [54]. La meilleure couverture lexicale peut être obtenue en retenant les mots les plus fréquents dans les données d'apprentissage, ou en ne prenant qu'un sous-ensemble des données (par exemple les données les plus récentes) [15, 37]. En moyenne, chaque mot hors vocabulaire est la cause de 1,5 à 2,0 erreurs [71]. Contrairement à une croyance largement répandue, un plus grand vocabulaire n'implique pas nécessairement un taux d'erreur plus élevé lorsqu'un modèle de langage adéquat est utilisé. Pour la plupart des systèmes les dictionnaires phonétiques utilisent des prononciations "standards" et ne représentent pas explicitement les allophones, laissant aux modèles acoustiques la représentation des variantes observées dans les données d'apprentissage. Plusieurs études ont été effectuées dans le but d'apprendre automatiquement ces prononciations, mais à notre connaissance ces approches quoique prometteuses n'ont pas encore permis d'améliorer significativement les performances des systèmes [82].

4. Modélisation linguistique

Les modèles de langage sont employés pour modéliser les régularités du langage naturel [78]. Les méthodes les plus utilisées sont fondées sur des statistiques n-grammes qui modélisent les contraintes syntaxiques et sémantiques en estimant la probabilité d'un mot dans un texte étant donnés les n-1 mots précédants. L'approche la plus commune pour "lisser" les statistiques des n-grammes rares est un mécanisme de repli utilisant des statistiques d'ordre inférieur lorsque les données d'apprentissage sont insuffisantes [16, 51]. Dans les systèmes actuels, les modèles de langage de type 3-gramme ou 4-gramme peuvent comprendre quelques dizaines de millions de paramètres. Le mécanisme de repli offre l'avantage supplémentaire que la taille du modèle peut être arbitrairement réduite en augmentant le nombre minimum d'observations requises pour inclure un n-gramme dans le modèle. Les modèles 2-gramme et 3-gramme sont les plus largement répandus. De petites améliorations ont été enregistrées avec l'utilisation de contexte plus large (4 ou 5-gramme) [6, 61, 97] ainsi qu'avec l'utilisation de modèles n-grammes de classe de mots [83].

Étant donné un corpus de textes (ou de transcriptions), il est relativement facile de construire un modèle n-gramme en comptant les occurrences de séquences de n mots [18]. Cependant, cela nécessite au préalable un travail important, tant pour la normalisation des textes avant qu'ils puissent être utilisés, que pour le choix du vocabulaire, la définition des mots, et le traitement des mots composés et des sigles. Fréquemment différentes sources de textes sont disponibles en quantités variables et doivent être combinées. Une solution à ce problème consiste à estimer un modèle par source puis de les interpoler. Les poids d'interpolation sont alors directement estimés sur des données de développement au moyen de l'algorithme EM.

5. Adaptation

Un des principaux défis en matière de reconnaissance de la parole est le développement de systèmes robustes, c.-à-d. qui conservent des performances élevées lorsque les conditions acoustiques de test et d'apprentissage sont différentes. Au niveau acoustique, deux classes de techniques pour augmenter la robustesse des systèmes peuvent être identifiées: les techniques de traitement du signal qui essayent de compenser la différence entre le test et l'apprentissage en modifiant le signal à décoder; et les techniques d'adaptation des modèles qui modifient les paramètres modèles pour les rendre plus représentatifs du signal observé.

Les approches fondées sur le traitement du signal comprennent les techniques de normalisation qui réduisent la variabilité du signal, augmentant les performances en conditions mal adaptées mais souvent avec une réduction des performances en conditions normales, et les techniques de compensation qui reposent sur un modèle du bruit et/ou un modèle de la parole. L'adaptation des modèles est une approche beaucoup plus puissante, en particulier quand le traitement du signal repose sur un modèle de la parole. Par conséquent quand les ressources en calcul ne sont pas considérées, l'adaptation des modèles est la solution de prédilection pour compenser les différences aussi minimes soient-elles.

Les techniques les plus généralement utilisées pour l'adaptation des modèles acoustiques, sont la composition de modèles [32, 33], l'adaptation bayésienne [38, 56, 88, 104], et des méthodes de transformation telles que la régression linéaire[59, 24]. La composition de modèles est essentiellement employée pour compenser des bruits additifs tandis que l'adaptation bayésienne et la régression linéaire sont des outils généraux qui peuvent être utilisés pour l'adaptation au locuteur et à l'environnement acoustique. La normalisation de la longueur du conduit vocal [3, 58, 91] est une autre technique qui a été proposée pour réaliser une certaine normalisation du signal vis-à-vis du locuteur.

Bien entendu l'adaptation peut concerner aussi bien le modèle de langage et le dictionnaire de prononciations que les modèles acoustiques. Diverses approches ont été proposées pour adapter le modèle de langage à partir des mots déjà reconnus dans le document à transcrire: un simple modèle de *cache* [49, 79], un modèle *trigger* [80], et un modèle de concordance de thème [87]. Le modèle de *cache* repose sur l'idée que les mots apparaissant dans un document qui vient d'être dicté ont une plus grande probabilité d'apparaître à nouveau. Pour les documents courts l'avantage de ce modèle est bien entendu très réduit. Le modèle *trigger* essaie de résoudre ce problème en augmentant les probabilités des mots qui apparaissent souvent dans les mêmes documents que les mots observés. Pour le modèle de concordance de thème, des mots-clés présents dans le discours traité sont utilisés pour rechercher des documents sur le même thème, documents à partir desquels des modèles de sous-langage sont élaborés puis utilisés pour redecoder le document courant. En dépit de l'intérêt croissant pour les modèles de langage adaptatifs, seules quelques améliorations minimes ont été obtenues par rapport à un modèle statique.

6. DÉCODEUR

L'un des défis posés par la reconnaissance à grands vocabulaire est la conception d'un algorithme de recherche efficace pour décoder l'énorme espace de recherche obtenu en combinant les modèles acoustique et linguistique. À proprement parler, le but du décodeur est de déterminer la suite de mots ayant la probabilité la plus élevée étant donnés le lexique et les modèles acoustique et linguistique. Dans la pratique, cependant, il est commun de rechercher la séquence d'états des MMC la plus probable, c.-à-d. le meilleur chemin dans un graphe (l'espace de recherche), où chaque nœud associe un état de MMC à une trame de signal. Puisqu'il est évidemment prohibitif de rechercher exhaustivement le meilleur chemin, des techniques ont été développées pour réduire le volume des calculs en limitant la recherche à une petite partie de l'espace total. L'approche la plus généralement utilisée pour de petites et moyennes tailles de vocabulaire est une recherche en faisceau trame-synchrone utilisant un algorithme de programmation dynamique [65]. Cette stratégie de base a été étendue pour traiter de grands vocabulaires en ajoutant des dispositifs tels que le *fast-match* [8, 39], les arbres phonétiques dépendants du mot précédant [66], la recherche avant-arrière [4], le réévaluation des N meilleures solutions [85], la recherche progressive [64] et le décodage dynamique en une passe [68]. Une alternative à la recherche trame-synchrone est une recherche asynchrone utilisant l'algorithme A* (*stack decodeur*) [7, 43, 74]. Les décodeurs dynamiques doivent faire appel à des techniques d'élagage très efficaces afin de prendre en compte toute l'information disponible en une seule passe. Ce type de décodeur est très attrayant pour des applications en temps réel. Cependant, beaucoup de systèmes en cours de développement utilisent les décodeurs à plusieurs passes pour réduire les besoins en calcul lorsque le décodage en temps réel n'est pas nécessaire [4, 36, 64, 76, 97].

Les techniques rapides de décodage sont essentielles pour le déploiement d'applications [84]. Pour des systèmes indépendants du locuteur avec des MMC multigaussiens, entre 30 et 50% du temps de décodage peut être utilisé pour évaluer les distributions gaussiennes. Ce temps peut être réduit d'une part en utilisant une méthode de calcul rapide pour les états des MMC, méthode qui bien entendu nécessite quelques approximations [10], et d'autre part en réduisant la taille des modèles avec des techniques de partage de paramètres, méthode qui a l'avantage de réduire également les besoins en mémoire. Un élagage agressif est généralement nécessaire pour effectuer le traitement en temps réel sur les plate-formes actuellement disponibles. C'est inévitablement une source d'erreurs de recherche, de sorte que de nombreuses techniques ont été proposées pour réduire ces erreurs de recherche et pour limiter leurs effets sur les performances des systèmes.

7. AU-DELÀ DES MOTS

En plus des mots prononcés, d'autres attributs peuvent être identifiés dans la signal audio. Cette information additionnelle peut être de nature linguistique (ponctuation, étiquettes sémantiques), ou de nature acoustique (identité du locuteur, environnement acoustique, tour de parole, mesure de confiance, ...).

En ce qui concerne les attributs de nature acoustique, les mêmes techniques de modélisation ont été appliquées avec succès à la reconnaissance du genre et de l'identité du locuteur, ainsi qu'à l'identification des conditions acoustiques. Pour le traitement d'un flux audio continu, il est avantageux de diviser les données en segments acoustiquement homogènes, et d'identifier et retirer les segments sans parole, puis de regrouper les segments de parole par locuteur. Ces informations peuvent être utilisées pour segmenter les transcriptions et ainsi faciliter leur indexation par un système de recherche documentaire.

Pour certaines applications, il peut être particulièrement utile d'estimer l'exactitude des mots et des phrases reconnus [14, 40, 89, 95, 96]. Pour les systèmes à grand vocabulaire, nous sommes essentiellement intéressés par une mesure de confiance au niveau du mot, le but étant d'estimer $\Pr(w_i|x)$ la probabilité *a posteriori* du i-ème mot du texte, ou alternativement $\Pr(w_i|x,\lambda)$ où λ représente les modèles du système. Une estimation de cette dernière probabilité peut être efficacement calculée en appliquant l'algorithme *forward-backward* à un graphe de mot produit par le système de reconnaissance en même temps que l'hypothèse [96]. Cette estimation reposant sur des modèles, bien entendu incorrects, il est commun d'utiliser d'autres caractéristiques du signal tels que les durées du mot et des phonèmes, le débit d'élocution, et le rapport signal/bruit pour obtenir une meilleure estimation de cette probabilité.

8. APPLICATIONS ET PERFORMANCES

La dictée de textes est l'application la plus évidente pour les systèmes de reconnaissance à grands vocabulaire. Elle a depuis 10 ans fait l'objet de développement de produits et il existe aujourd'hui des logiciels peu coûteux disponibles pour une variété de langages et de plateformes matérielles. Sans doute la caractéristique la plus notable de ce type d'application est que la parole à traiter est produite dans le but explicite d'être transcrite par une machine. Cette application a été largement utilisée pour mesurer les progrès en matière de RAP, car il est facile d'évaluer les résultats en comparant la transcription automatique à une transcription de référence. La métrique généralement utilisée est le taux d'erreur sur les mots défini comme suit : *%erreur = %substitutions + %insertions + %élisions*. Sur des données du corpus NAB du LDC (*North American Business News*, textes lus, micro casque), l'état de l'art pour des systèmes indépendants du locuteur se situe autour de 7% d'erreurs. Les mêmes données enregistrées avec un microphone de table dans un environnement bruité (55dBA, S/B de 15dB), le taux d'erreur est environ 14% avec compensation du bruit [71, 72]. Le taux d'erreur pour la dictée spontanée d'articles financiers est de l'ordre de 14% et est supérieur à 20% pour des textes lus au téléphone. En français, sur le corpus BREF de textes lus du journal Le Monde, le taux d'erreur est d'environ 10% [25] (pour des travaux français sur ce problème cf. [2, 13, 31]).

Le second domaine d'applications concerne la transcription et l'indexation des données audio en général, telles que des émissions de radio et télévision, des téléconférences, ou tout autre document audio susceptible d'être indexé [12, 50, 52, 69, 73]. Plusieurs caractéristiques de ce type de données peuvent être identifiées. D'abord, on peut considérer qu'il s'agit de données "trouvées", qui ne sont pas produites dans le but d'être traitées par une machine. En second lieu, il s'agit de flux audio continus, avec de nombreux changements de locuteurs, sans aucune segmentation à priori. Troisièmement, la prise de son et l'environnement acoustique sont beaucoup moins contrôlés que pour les systèmes de dictée. Sur des documents d'information (radio et TV), le taux d'erreur moyen est de l'ordre de 20% pour l'anglais-américain et environ de 25% pour le français et l'allemand. Une section spéciale de la revue CACM a été récemment consacrée à ce sujet [63]. Sur la tâche de DARPA Hub5 [42] adressant la transcription de la parole conversationnelle au téléphone, le taux d'erreur se situe autour de 40% [102].

La troisième classe d'application est celle des systèmes de dialogue [20]. La plupart de ces systèmes visent à offrir un accès à des bases de données. Il y a de plus en plus de systèmes opérationnels mais ils emploient généralement des stratégies de dialogue beaucoup plus contraintes que les prototypes de laboratoire dits à initiative partagée. L'éventail des taux d'erreur de reconnaissance qui ont été publiés pour ces systèmes s'étend de 5% pour des tâches simples (horaires d'avions) avec micro casque à plus de 25% pour des serveurs téléphoniques.

9. DÉFIS ET PERSPECTIVES

La reconnaissance de la parole est loin d'être un problème résolu, comme cela est démontré par la grande différence entre les performances de la machine et celle de l'auditeur humain [29, 92, 60]. Pour combler cette différence nous devons sans aucun doute améliorer nos modèles à tous les niveaux: acoustique, lexical, syntaxe et sémantique.

Pour les systèmes indépendants du locuteur, il est bien connu qu'il peut y avoir une énorme différence (jusqu'à un rapport 20) entre les taux d'erreur de deux locuteurs [30]. Ceci peut être attribué à une variété de facteurs liés au locuteur et à sa vitesse d'élocution [71]. L'adaptation des modèles acoustiques permet de compenser en partie cette différence, mais nécessite au moins quelques minutes de signal pour être vraiment efficace, ce qui limite son champ d'application. Le développment de techniques d'adaptation plus efficaces et plus rapides qui prennent mieux en compte les corrélations entre les paramètres des modèles est donc une nécessité. La réduction de cette différence doit sans doute aussi passer par l'adaptation du lexique de prononciations, en généralisant les variantes observées sur les données déjà produites par le locuteur. Une personne qui prononce un mot d'une façon donnée est susceptible de prononcer les mots semblables de la même manière. Pour les règles de coarticulation entre mots, différents locuteurs appliquent différentes règles phonologiques, et bien que ces règles soient habituellement systématiques pour un même locuteur, à notre connaissance aucun système ne sait tirer parti de cette information.

Côté modélisation linguistique, les techniques explorées pour effectuer les accords à long terme n'ont pas encore été couronnées de succès. Ces techniques seraient particulièrement utile pour traiter les langues fortement flexionnelles pour lesquelles les modèles n-grammes ne sont clairement pas la solution optimale. L'adaptation des modèles linguistiques est un défi pour les systèmes de transcription de documents d'information radio et TV, où il est particulièrement important de maintenir les modèles à jour. De nouveaux thèmes peuvent apparaître soudainement, et demeurer dans l'actualité pendant un temps très variable. L'existence de sources de données contemporaines, tels que les journaux électroniques disponibles sur Internet, devrait nous permettre de mettre à jour automatiquement les modèles de langage [52].

Le développement de systèmes indépendants de l'application est un autre défi majeur. A partir d'un grand corpus de parole transcrite, il est possible de développer des modèles acoustiques pour une variété d'applications, il n'en est pas de même pour les modèles de langage où une bonne couverture du domaine de l'application est essentielle. Avec la technologie actuelle, le portage d'un système vers une nouvelle application ou une autre langue nécessite l'existence de quantités suffisantes de données transcrites. Le développement de techniques d'apprentissage nécessitant peu de supervision est donc un axe de recherche à explorer.

10. CONCLUSION

En dépit des nombreuses avancées de cette dernière décennie, et de la généralisation des systèmes de dictée de textes, la reconnaissance de la parole est loin d'être un problème résolu. Alors qu'il est clair que nos modèles ont besoin d'être améliorés en particulier pour la parole conversationnelle, nous ne savons pas quel est le chaînon le

plus faible entre le modèle acoustique, le modèle de langage et le dictionnaire de prononciations.

Il apparaît cependant qu'une vaste gamme d'applications est maintenant rendue accessible. Les deux domaines les plus prometteurs concernent les serveurs d'informations, et les systèmes d'indexation automatique de documents audio. Les premières expériences en recherche documentaire dans des documents audio ont conduit à des résultats comparables en utilisant des transcriptions manuelles et automatiques. L'énorme quantité d'information diffusée quotidiennement sous formes audio et audiovisuelle nous permet de mesurer l'intérêt de ce résultat.

BIBLIOGRAPHIE

[1] D. Abberley, D. Kirby, S. Renals et T. Robinson, "The THISL Broadcast News Retrieval System," *Proc. ESCA ETRW on Accessing Information in Spoken Audio*, pp. 14-19, Cambridge, U.K., avril 1999.

[2] G. Adda, M. Adda-Decker, J.L. Gauvain, et L. Lamel, "Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97", *JST FRANCIL*, Avignon, avril 1997.

[3] A. Andreoum T. Kamm et J. Cohen, "Experiments in Vocal Tract Normalisation", *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[4] S. Austin, R. Schwartz et P. Placeway, "The Forward-Backward Search Strategy for Real-Time Speech Recognition," *Proc. IEEE ICASSP-91* pp. 697-700, Toronto, mai 1991.

[5] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, et H.F. Silverman,"Preliminary results on the performance of a system for the automatic recognition of continuous speech," *Proc. IEEE ICASSP-76*, Philadelphia, PA, avril 1976.

[6] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan et S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognizer for the ARPA NAB News Task," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 121-126, Austin, TX, janvier 1995.

[7] L.R. Bahl, F. Jelinek et R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-5**(2), pp. 179-190, mars 1983.

[8] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo et M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *Proc. IEEE ICASSP-92*, CA, **1**, pp. 17-21, San Francisco, CA, mars 1992.

[9] J. Baker, J. Baker, P. Bamberg, K. Bishop, L. Gillick, V. Helman, Z. Huang, Y. Ito, S. Lowe, B. Peskin, R. Roth et F. Scattone, "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc. DARPA Speech and Natural Language Workshop*, pp. 387-392, Harriman, NY, février 1992.

[10] E. Bocchieri, "Vector quantization for efficient computation of continuous density likelihoods," *Proc. IEEE ICASSP-93*, **2**, pp. 692-695, Minneapolis, MN, mai 1993.

[11] H. Bourlard et N. Morgan, "Continuous Speech Recognition by Connectionist Statistical Methods," *IEEE Trans. on Neural Networks*, **4**(6), pp. 893-909, 1994.

[12] F. Brugnara, M. Cettolo, M. Federico et D. Giuliani, "A Baseline for the Transcription of Italian Broadcast News," *Proc. IEEE ICASSP-00*, Istanbul, Turkey, juin 2000.

[13] M.J. Caraty, C. Montacié et F. Lefèvre, "Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation System", *Eurospeech*, Rhodes, pp. 2691-2694, 1997.

[14] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition", *Proc. ESCA Eurospeech'97*, pp. 815-818, Rhodes, Greece, septembre 1997.

[15] L. Chase, R. Rosenberg, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide et C. Lu, "Improvements in Language, Lexical and Phonetic Modeling in Sphinx-II," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 60-65, Austin, TX, janvier 1995.

[16] S.F. Chen et J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer, Speech and Language*, **13**(4), pp. 359-394, octobre, 1999.

[17] Y.L. Chow, R. Schwartz, S. Roukos, O. Kimball, P. Price, F. Kubala, M.O. Dunham, M. Krasner et J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *Proc. IEEE ICASSP-86*, **3**, pp. 1593-1596, Tokyo, Japan, avril 1986.

[18] P. Clarkson et R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge Toolkit," *Proc. ESCA EuroSpeech'97*, pp. 2707-2710, Rhodes, Greece, septembre 1997.

[19] S. Davis et P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **28**(4), pp. 357-366, 1980.

[20] R. De Mori, "Spoken Dialogues with Computers," Academic Press, 1998.

[21] N. Deshmukh, A. Ganapathiraju, R.J. Duncan et J. Picone, "Human Speech Recognition Performance on the 1995 CSR Hub-3 Corpus" *Proc. ARPA Speech Recognition Workshop*, pp. 129-134, Harriman, NY, février 1996.

[22] V. Digalakis et H. Murveit, "Genones: Optimization the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," *Proc. IEEE ICASSP-94*, **1**, pp. 537-540, Adelaide, Australia, avril 1994.

[23] V. Digalakis, M. Weintraub, A. Sankar, H. Franco, L. Neumeyer et H. Murveit, "Continuous Speech Dictation on ARPA's North American Business News Domain," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 88-93, janvier 1995.

[24] V. Digalakis, D. Rtichev et L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Trans. on Speech and Audio*, **3**(5), 357-366, septembre 1995.

[25] J.M. Dolmazon, F. Bimbot, G. Adda, M. El Beze, J.C. Caerou, J. Zeiliger et M.A Decker, "ARC B1 - Organisation de la 1e campagne AUPELF pour l'évaluation des systèmes de dictée vocale," *1ères JST FRANCIL*, Avignon, avril 1997.

[26] J. Dreyfus-Graf, "Sonograph and Sound Mechanics," *J. Acoust. Soc. America*, **22**, pp. 731, 1949.

[27] H. Dudley et S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," *J. Acoust. Soc. America*, **30**, pp. 721, 1958.

[28] C. Dugast, R. Kneser, X. Aubert, S. Ortmanns, K. Beulen et H. Ney, "Continuous Speech Recognition Tests and Results for the NAB'94 Corpus," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 156-161, janvier 1995.

[29] W.J. Ebel et J. Picone, "Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus" *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 53-59, Austin, TX, janvier 1995.

[30] W. Fisher, "Factors Affecting Recognition Error Rate," *Proc. ARPA Speech Recognition Workshop*, pp. 47-52, Harriman, NY, février 1996.

[31] D. Fohr, J.P. Haton, J.F. Mari, K. Smaïli et I. Zitouni, "MAUD : un prototype de machine à dicter vocale", *1ères JST FRANCIL*, Avignon, avril 1997.

[32] M.J.F. Gales et S.J. Young, "An improved approach to hidden Markov model decomposition of speech and noise," *Proc. IEEE ICASSP-92*, pp. 233-236, San Francisco, CA, mars 1992.

[33] M.J.F. Gales et S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), pp. 289-307, octobre 1995.

[34] J.L. Gauvain et L. Lamel, "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications," *Institute of Electronics, Information and Communication Engineers*, J79-D-II:2005–2021, decembre 1996.

[35] J.L. Gauvain, L.F. Lamel, G. Adda et M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, octobre 1994.

[36] J.L. Gauvain, L.F. Lamel, G. Adda et M. Adda-Decker, "The LIMSI Nov93 WSJ System," *Proc. ARPA Spoken Language Technology Workshop*, pp. 125-128, Princeton, NJ, mars 1994.

[37] J.L. Gauvain, L.F. Lamel et M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proc. IEEE ICASSP-95*, pp. 65-68, Detroit, MI, mai 1995.

[38] J.L. Gauvain and C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Processing*, **2**(2), pp. 291-298, April 1994.

[39] L. Gillick et R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 170-172, Hidden Valley, PA, juin 1990.

[40] L. Gillick, Y. Ito et J. Young, "A Probabilistic Approach to Confidence Measure Estimation and Evaluation", *Proc. IEEE ICASSP-97*, pp. 879-882, Munich, Germany, avril 1997.

[41] J.R. Glass, T.J. Hazen et I. L. Hetherington, "Real-time Telephone-based Speech Recognition in the Jupiter Domain," *Proc. IEEE ICASSP-99*, **1**, pp. 61-64, Phoenix, AZ, mars 1999.

[42] J. Godfrey, E. Holliman et J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. IEEE ICASSP-92*, pp. 517-520, San Francisco, CA, mars 1992.

[43] P.S. Gopalakrishnan, L.R. Bahl et R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 572-575, Detroit, MI, mai 1995.

[44] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, **87**(4), pp. 1738-1752, 1990.

[45] M.M. Hochberg, S.J. Renals, A.J. Robinson et D. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system," in *Proc. ICSLP'94*, pp. 1499-1502, Yokohama, Japan, septembre 1994.

[46] M. Hwang et X. Huang, "Subphonetic Modeling with Markov States - Senone," *Proc. IEEE ICASSP-92*, San Francisco, CA, **1**, pp. 33-36, mars 1992.

[47] X. Huang, F. Alleva, M.Y. Hwang et R. Rosenfeld, "An Overview of the SPHINX-II Speech Recognition System," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 81-86, mars 1993.

[48] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, avril 1976.

[49] F. Jelinek, B. Merialdo, S. Roukos et M. Strauss, "A Dynamic Language Model for Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 293-295, Pacific Grove, CA, février 1991.

[50] F. deJong, J.L. Gauvain, J. deb Hartog, K. Netter, "OLIVE: Speech Based Video Retrieval," *Proc. CBMI'99*, Toulouse, octobre 1999.

[51] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-35**(3), pp. 400-401, mars 1987.

[52] T. Kemp et A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. ESCA Eurospeech'99*, Budapest, Hungary, **6** 2725-2728, septembre 1999.

[53] D. Kershaw, A.J. Robinson et S.J. Renals, "The 1995 Abbot hybrid connectionist-HMM large-vocabulary recognition system," *Proc. ARPA Speech Recognition Workshop*, pp. 93-98, Harriman, NY, février 1996.

[54] L.F. Lamel et G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition," *Proc. ICSLP'96*, **1**, pp. 6-9, Philadelphia, PA, octobre 1996.

[55] L. Lamel, G. Adda et M. Adda-Decker, "Les lexiques de prononciation dans les systèmes de reconnaissance de la parole," *Proc. Séminaire GDR-PRC CHM Lexique et communication parlée*, pp. 1-10, Toulouse, octobre 1996.

[56] C.-H. Lee et Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," to appear in *Proc. of the IEEE*, special issue, 2000.

[57] K.-F. Lee, *Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system*, PhD Thesis, Carnegie Mellon University, 1988.

[58] L. Lee et R.C. Rose, "Speaker Normalisation Using Efficient Frequency Warping Procedures", *Proc. IEEE ICASSP-96*, **1**, pp. 353-356, Atlanta, GA, mai 1996.

[59] C.J. Leggetter et P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, **9**, pp. 171-185, 1995.

[60] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, **22**(1), pp. 1-15, 1997.

[61] A. Ljolje, M.D. Riley, D.M. Hindle et F. Pereira, "The AT&T 60,000 Word Speech-To-Text System," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 162-165, Austin, TX, janvier 1995.

[62] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui et K. Shirai, "Large-Vocabulary Continuous Speech Recognition using the Japanese Business Newspaper (Nikkei) Task," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, février 1996.

[63] M. Maybury (ed.), "News on Demand," Special section in the *Communications of the ACM* **43**(2), février 2000.

[64] H. Murveit, J. Butzberger, V. Digalakis et M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proc. IEEE ICASSP-93*, **II**, pp. 319-322, Minneapolis, MN, avril 1993.

[65] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-32**(2), pp. 263-271, avril 1984.

[66] H. Ney, R. Haeb-Umbach, B.H. Tran et M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE ICASSP-92*, **I**, pp. 9-12, San Francisco, CA, mars 1992.

[67] L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos et Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 77-81, janvier 1995.

[68] J.J. Odell, V. Valtchev, P.C. Woodland et S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. ARPA Human Language Technology Workshop*, pp. 405-410, Princeton, NJ, mars 1994.

[69] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki et Z.P. Zeang, "Recent Advances in Japanese Broadcast News Transcription," *Proc. ESCA Eurospeech'99*, **2**, pp. 671-674, Budapest, Hungary, septembre 1999.

[70] M. Ostendorf, A. Kannan, O. Kimball et J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proc. ARPA Workshop on Continuous Speech Recognition*, pp. 53-58, Stanford, CA, septembre 1992.

[71] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin et M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 5-36, Austin, TX, janvier 1995.

[72] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, A.F. Martin et M.A. Przybocki, "1995 Hub-3 Multiple Microphone Corpus Benchmark Tests," *Proc. ARPA Speech Recognition Workshop*, pp. 27-46, Harriman, NY, février 1996.

[73] D.S. Pallett, A.F. Martin et M.A. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," *Proc. DARPA Broadcast News Workshop* pp. 5-12, Herndon, VA, février 1999.

[74] D.B. Paul, "An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model," *Proc. IEEE ICASSP-92*, pp. 405-409, San Francisco, CA, mars 1992.

[75] D.B. Paul, "New Developments in the Lincoln Stack-Decoder Based Large Vocabulary CSR System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 143-147, janvier 1995.

[76] F. Richardson, M. Ostendorf et J.R. Rohlicek, "Lattice-Based Search Strategies for Large Vocabulary Recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 576-579, Detroit, MI, 1995.

[77] I. Rogina et A. Waibel, "The JANUS Speech Recognizer," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 166-169, janvier 1995.

[78] R. Rosenfeld, "Adaptive Statistical Language Modeling," to appear in *Proc. of the IEEE*, special issue, 2000.

[79] R. Rosenfeld, *Adaptive Statistical Language Modeling*, PhD Thesis, Carnegie Mellon University, 1994. (also *Tech. rep. CMU-CS-94-138*)

[80] R. Rosenfeld et X. Huang, "Improvements in Stochastic Language Modeling," *Proc. DARPA Workshop on Speech and Natural Language*, pp. 107-111, Harriman, NY, février 1992.

[81] R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann et J. Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer, *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 116-120, janvier 1995.

[82] M.D. Riley, W. Byrne, M. Finke, S. Khudanpu, A. Ljojle, J. McDonough, H. Nock, M. Saraclar, C. Wooters et G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Automatic Speech and Speaker Recognition, Speech Communication* **29**(2-4), pp. 209-224, novembre 1999.

[83] A. Sankar, A. Stolke, T. Chung, L. Neumeyer, M. Weintraub, H. Franco et F. Beaufays, "Noise-Resistant Feature Extraction and Model Training for Robust Speech Recognition," *Proc. ARPA Speech*

Recognition Workshop, pp. 117-122, Harriman, NY, février 1996.

[84] M. Schuster, "Memory-efficient LVCSR search using a one-pass stack decoder," *Computer, Speech and Language*, **14**(1), pp. 47-77, janvier 2000.

[85] R. Schwartz, S. Austin, F. Kubala et J. Makhoul,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *Proc. IEEE ICASSP-92*, **I**, pp. 1-4, San Francisco, CA, mars 1992.

[86] R. Schwartz, Y. Chow, S. Roucos, M. Krasner et J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," *Proc. IEEE ICASSP-84*, **3**, pp. 35.6.1-35.6.4, San Diego, CA, mars 1984.

[87] S. Sekine et R. Grishman, "NYU Language Modeling Experiments for the 1995 CSR Evaluation," *Proc. ARPA Speech Recognition Workshop*, pp. 123-128, Harriman, NY, février 1996.

[88] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *Proc. IEEE ICASSP-95*, pp. 697-700, Detroit, MI, mai 1995.

[89] M. Siu et H. Gish, "Evaluation of word confidence for speech recognition systems", *Computer Speecd & Language*, **13**(4), pp. 299-318, octobre 1999.

[90] S. Takahashi et S. Sagayama, "Four-level Tied Structure for Efficient Representation of Acoustic Modeling," *Proc. IEEE ICASSP-95*, pp. 520-523, Detroit, MI, mai 1995.

[91] L.F. Uebel et P.C. Woodland, "An Investigation into Vocal Tract Length Normalisation", *Proc. ESCA Eurospeech'99*, pp. 2527-2530, Budapest, Hungary, septembre 1999.

[92] D.A. van Leeuwen, L.G. van den Berg, H.J.M. Steeneken, "Human Benchmarks for Speaker Independent Large Vocabulary Recognition Performance," *Proc. ESCA Eurospeech'95*, pp. 1461-1464, Madrid, Spain, septembre 1995.

[93] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibnernetika*, **4**, p. 81, 1968.

[94] T.K. Vintsyuk, "Elements-wise recognition of continuous speech composed of words from a specified dictionary," *Cybernetics*, **7**, pp. 133-143, mars-avril 1971.

[95] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig et A. Stolcke, "Neural-Network based Measures of Confidence for Word Recognition," *Proc. ICASSP-97*, pp. 887-890, Munich, Germnay, avril 1997.

[96] F. Wessel, K. Macherey et R. Schlüter, "Using word probabilities as confidence measures," *Proc. IEEE ICASSP-98*, pp. 225-228, Seattle, WA, mai 1998.

[97] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev et S.J. Young, "The development of the 1994 HTK large vocabulary speech recognition system," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 104-109, Austin, TX, janvier 1995.

[98] P.C. Woodland, M.J.F. Gales, D. Pye et V. Valtchev, "The HTK large vocabulary recognition system for the 1995 ARPA H3 task," *Proc. ARPA Speech Recognition Workshop*, pp. 99-104, Harriman, NY, février 1996.

[99] S.J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognisers," *Proc. IEEE ICASSP-92*, San Francisco, CA, pp. 569-572, mars 1992.

[100] S.J. Young, "A Review of Large-Vocabulary Continuous Speech Recognition," *IEEE Signal Processing Magazine*, **13**(5), pp. 45-57, septembre 1996.

[101] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken A.J. Robinson et P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project," *Computer Speech and Language*, **11**(1):73-89, janvier 1997.

[102] S.J. Young et L. Chase, "Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes," *Computer Speech and Language*, **12**(4), pp. 263-279, octobre 1998.

[103] S.J. Young et P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. ESCA Eurospeech'93*, **3**, pp. 2203-2206, Berlin, Germany, septembre 1993.

[104] G. Zavaliagkos, R. Schwartz et J. McDonough, "Maximum *a Posteriori* Adaptation for Large Scale HMM Recognizers," *Proc. IEEE ICASSP-95*, pp. 725-728, Detroit, MI, mai 1995.

[105] V. Zue. J. Glass, M. Phillips et S. Seneff, "The MIT SUMMIT Speech Recognition System: A Progress Report", *Proc. DARPA Speech and Natural Language Workshop*, pp. 179-189, Philadelphia, PA, février 1989.

Imageries fonctionnelles cérébrales : vers une physiologie de la cognition humaine

Jean-François Démonet

INSERM U 455 , Hôpital Purpan, 31059 Toulouse cedex 03
Tel : 05 61 77 95 00 Fax : 05 61 49 95 24 Email : demonet@purpan.inserm.fr

ABSTRACT

The recent progress of functional imaging techniques has renewed our understanding of mind / brain relationships that represent the fundamental topic of Neuropsychology. Without requiring observations in brain-damaged patients, functional neuro-imaging provide informations on the neural substrates of cognitive processes such as language, memory, or attention. These functions correspond to large-scale neural ensembles distributed throughout the entire brain, rapidly evolving over very short periods of time, and characterised by complex, non-linear dynamics. Because of this complexity, (i) the use of functional neuro-imaging requires careful consideration of a number of methodological issues on the relationships between cognitive processes under study and changes in functional signals that are recorded , and (ii) only the combination of these techniques, providing spatial resolution on the one hand (PET, fMRI) and temporal resolution on the other (EEG , MEG) could improve our knowledge of the spatial-temporal dynamics of such neural ensembles.

INTRODUCTION

Pendant les cent dernières années, un cadre théorique unique a sous-tendu la problématique qui, trouvant ses racines dans l'Antiquité (Messerli, 1993), est devenue celle de la neuropsychologie naissante. Celle-ci a pour objet les relations entre structure et fonction, c'est à dire entre cerveau et esprit. Ce cadre théorique de référence est le **paradigme des lésions** que le milieu neurologique français désigne en général sous le vocable "méthode anatomo-clinique". Fondée notamment par Broca, Marie et Déjerine, elle consiste à caractériser le plus précisément possible l'anatomie des lésions et les symptômes neuropsychologiques qui en découlent, puis à rechercher les correspondances entre ces deux ensembles de données.

L'avènement des techniques d'imagerie fonctionnelle cérébrale et leurs progrès croissants en termes de sensibilité et de résolution ont créé, au cours des dernières années, un contexte favorable à l'établissement d'un **nouveau cadre expérimental** (Jeannerod, 1996). Ce paradigme - que l'on pourrait dire "des fonctions" ou "des variations" par opposition à celui des lésions - renouvelle en effet l'approche des relations entre cerveau et esprit. Il complète les connaissances antérieures en ce qu'il apporte des informations acquises **indépendamment** du paradigme des lésions.

L'imagerie fonctionnelle du cerveau consiste en la mesure d'indices externes, généralement très indirects, reflètant l'activité métabolique au sein de populations de neurones cérébraux. Considérées à l'échelon du neurone ou de la synapse, avec des constantes de l'ordre du micron pour l'espace et de la milliseconde pour le temps, ces méthodes sont fort imprécises. Cette imprécision ne doit cependant pas être considérée de façon seulement négative. En effet il est bien clair que le niveau de description pertinent en ce qui concerne les corrélats neuronaux des fonctions cognitives n'est pas celui de la synapse mais correspond vraisemblablement à celui d'ensembles neuronaux inter-connectés et très largement **distribués** dans l'espace cérébral. Plus ennuyeux est le fait qu'aucune des méthodes d'imagerie actuelles ne combine une résolution satisfaisante dans les deux domaines - l'espace anatomique de ces ensembles neuronaux et leur constante de temps - c'est à dire de l'ordre du millimètre et de la milliseconde. Les techniques tomographiques (PET et IRM) ont pour elles une résolution anatomique correcte mais sont défavorisées en ce qui concerne la résolution temporelle. Cette dernière est excellente lorsque l'on utilise les techniques électro- ou magnéto-encéphalographiques qui sont, réciproquement, limitées dans leur résolution spatiale. Sur le plan technique, les progrès à attendre sont donc des possibilités de combinaison ou de fusion des données issues de plusieurs techniques, exploitant ainsi les avantages de chacune.

REPRESENTATIONS DE L'ACTIVITE CEREBRALE

Outre le caractère distribué dans l'espace cérébral des corrélats cérébraux des fonctions cognitives, un second aspect de leur complexité consiste dans le caractère **dynamique** des phénomènes qu'il s'agit d'identifier au sein de ces ensembles neuronaux. Les états mentaux, se reflétant dans les cartes cérébrales que nous nous efforçons de dresser, sont, par essence, labiles, dans une gamme de temps inférieure à la seconde. **La physiologie de la cognition doit adapter sa méthodologie à l'instabilité fondamentale des phénomènes qu'elle entend saisir.** Le nouveau paradigme expérimental a donc bien pour objet les **variations de phénomènes neurophysiologiques.** Les variables manipulées en imagerie fonctionnelle consistent essentiellement en des valeurs relatives appréciant le **changement de tel indice de fonctionnement neuronal par rapport à une autre mesure de l'activité cérébrale, effectuée dans une circonstance différente.** D'emblée, se trouve donc posé le

problème d'un niveau de référence par rapport auquel ces changements pourraient être appréciés. En fait, la référence elle-même semble souvent n'être que relative, définie seulement par le contexte expérimental propre à chaque étude d'imagerie. Ainsi, il ne paraît pas exister, pour l'étude des variations d'activité cérébrale, de niveau de base absolu qui définirait un état de repos cérébral bien improbable en ce que le « vrai repos » du cerveau serait proche du "repos éternel" ! **La physiologie de la cognition se doit donc de capter la dynamique d'activité cérébrale en tentant de l'influencer quelque peu, ... à défaut de pouvoir la maîtriser complètement.** La méthode la plus utilisée dans le cadre du paradigme des variations est appelée méthode d'"**activation cérébrale.** » Cette méthode suppose que la **manipulation expérimentale de fonctions neurophysiologiques** (stimulation sensorielle, mise en jeu de processus cognitifs ou moteurs, etc...) **au cours de la mesure** d'activité cérébrale, est susceptible de provoquer des variations détectables d'activité, dans des régions anatomiques et/ou des domaines temporels, représentatifs des fonctions suscitées, transitoirement, par l'expérience. Il faut noter le caractère trompeur du terme "activation", dans la mesure où il suggère uniquement l'existence d'augmentations d'activité liées à l'exercice des fonctions physiologiques. En fait, une bonne part des résultats, très largement passée sous silence dans la littérature, consiste en l'observation de **diminutions** d'activité dans certaines régions cérébrales lorsqu'une fonction s'effectue (Figure 1).

La mise en relation d'une fonction physiologique avec des variations locales du débit sanguin, reflet indirect de l'activité neuronale correspondante, semble remonter aux travaux de Roy et Sherrington (1890). Les méthodes d'imagerie du débit sanguin cérébral dont est issue la **tomographie par émission de positons (PET)**, sont basées sur l'administration d'un bolus de radio-traceur de haute énergie et brève durée de vie et sur l'évolution de la radio-activité mesurée au niveau céphalique pendant **plusieurs dizaines de secondes.** Dès les premières études isotopiques du débit sanguin cérébral régional chez l'homme (Lassen et Ingvar, 1961), on a obtenu une **validation empirique** de la méthode d'"activation cérébrale". En effet, quelles que soient les réserves faites plus haut quant au problème de la mesure de référence, des effets très clairs ont pu être détectés, consistant en des augmentations de débit dans les régions correspondant aux cortex sensoriels pour une mesure faite pendant la stimulation sensorielle par rapport à une mesure de référence au cours de laquelle aucune stimulation n'est, volontairement, appliquée. Des phénomènes similaires furent constatés au niveau des cortex moteurs pendant que des sujets effectuaient un mouvement, par rapport à une situation de repos, sans mouvement volontaire. **Ces augmentations de débit dans les cortex primaires sensoriels ou moteurs (se situant donc en somme à l' »entrée » ou à la « sortie » du système) sont importantes, de l'ordre de 20 % et parfois plus.**

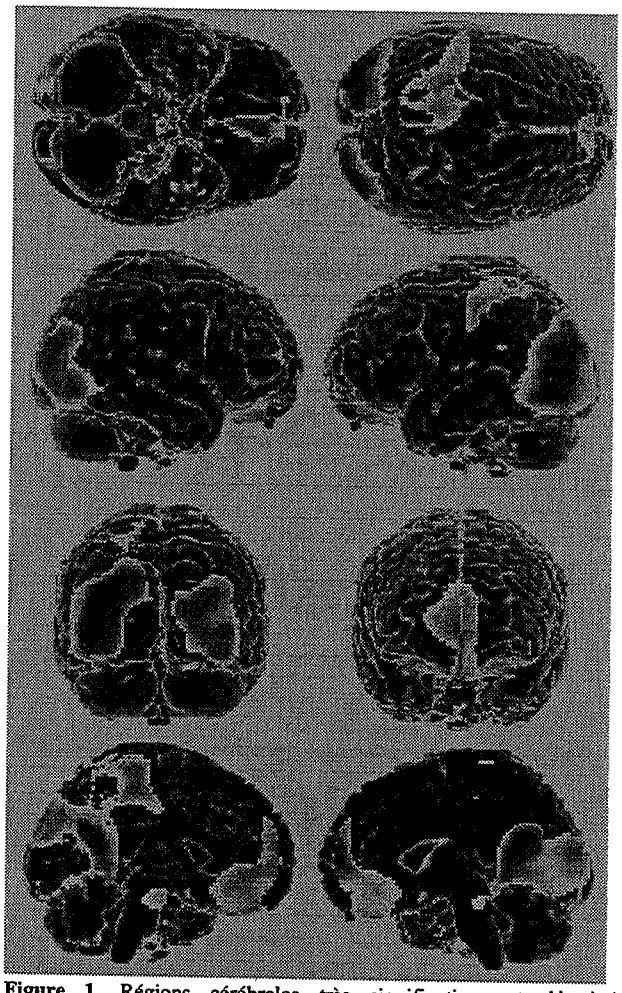


Figure 1. Régions cérébrales très significativement déactivées (p<.0001 après correction pour des comparaisons statistiques multiples) dans un groupe de 6 sujets normaux. Les pixels significatifs sont projetés sur des représentations surfaciques tri-dimensionnelles d'un cerveau « standard » (cf. logiciel SPM96, Frackowiak & Friston). Les valeurs de débit sanguin cérébral estimées sont beaucoup plus fortes dans ces régions au cours de la condition « de repos » par rapport aux conditions « actives » (tâches linguistiques de jugement de rime). Ce profil de régions déactivées comprenant les régions corticales pariéto-occipitales, la région cingulaire postérieure et les régions fronto-polaires est très régulièrement retrouvé dans nombre de travaux aussi bien en T.E.P. qu'en IRMf (voir par exemple Shulman et al., 1998).

Bien que les variations de débit au niveau des cortex associatifs (fonctionnellement « situés » entre entrées et sorties) soient beaucoup plus faibles (de l'ordre de 5 % voire moins), on a voulu, très vite, appliquer cette logique élémentaire d'activation (stimulus/réponse) aux fonctions cognitives. La conception des expériences d'activation cognitive fut donc basée sur une logique additive supposant une **correspondance quasi terme-à-terme** entre les composantes cognitives des tâches servant à l'expérience d'activation et leurs éventuels corrélats cérébraux, mis en évidence par comparaison entre deux ou plusieurs mesures. (Figure 2). Cette conception est basée sur plusieurs présupposés: (a) les tâches (et leurs conséquences cérébrales) se distinguent entre elles **en termes binaires de tout ou rien** (présence ou absence de tel ou tel composant cognitif dans une tâche donnée), (b) chacun de ses composants induit un effet nettement

mesurable en termes d'activité cérébrale et ce d'une manière **indépendante** par rapport aux autres composants cognitifs présents dans l'expérience et à leurs effets d'activation éventuels. Ainsi, les tâches d'activation peuvent être vues comme la somme de composants indépendants. Les comparaisons entre tâches peuvent être donc être conçues de manière **hiérarchique**, en fonction d'une complexité croissante, obtenue par **addition successive de composants cognitifs** (en anglais cette conception est désignée par le terme "pure insertion" pour exprimer l'indépendance des composants entre eux).

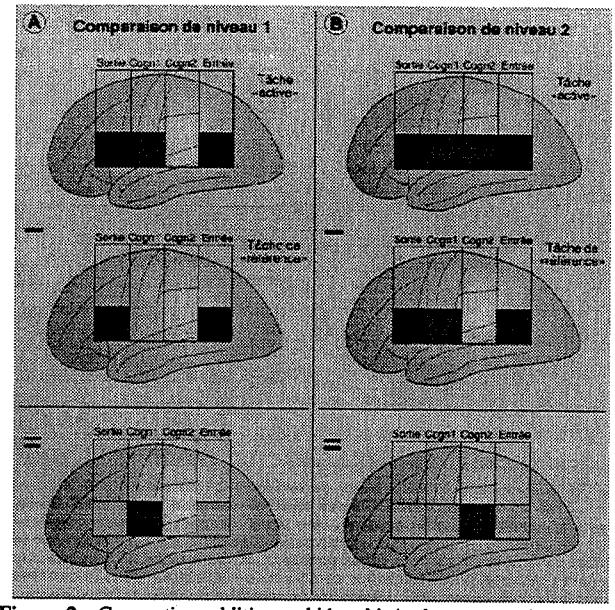


Figure 2. Conception additive et hiérarchisée des comparaisons entre tâches. Un article de Petersen et al. (1988) qui a marqué les débuts des recherches modernes en matière d'imagerie fonctionnelle des fonctions cognitives décrivait des résultats d'activation lié au traitement de mots isolés perçus visuellement ou auditivement. Cette conception strictement additive et hiérarchisée (cf. texte) des tâches d'activation entre elles (et de leurs équivalents neurofonctionnels) conduisait les auteurs à décrire comme seul corrélat cérébral aux processus d'association sémantique (tâche de génération de verbes) l'activation du cortex préfrontal inférieur gauche (aire 47) alors que des travaux ultérieurs ont montré l'implication de tout un réseau impliquant aussi des régions temporales et pariétales (cf. Vandenberghe et al., 1996). Selon la conception additive et hiérarchisée, l'addition progressive de nouvelles composantes permet de créer une hiérarchie de complexité croissante entre différentes tâches d'activation. La « soustraction » des composantes et de leurs équivalents neurofonctionnels permettent d'isoler successivement le corrélat cérébral de chaque composante. En A, par exemple, la comparaison de niveau 1 pourrait impliquer une tâche de base avec stimulation verbale auditive et réponse stéréotypée et une tâche expérimentale de répétition des mots perçus auditivement, aboutissant à la localisation des processus de transposition audio-phonatoire. En B, dans la comparaison de niveau 2, la répétition est considérée à son tour comme tâche de base pour une tâche plus complexe, par exemple évoquer un verbe sémantiquement associé à un nom d'objet perçu auditivement. Le résultat net correspondrait alors à la localisation des processus sémantiques, tout autre effet lié à des processus de « niveau inférieur » (perception auditive, formulation verbale, etc ...) étant éliminé par le fait qu'ils sont également présents dans les deux tâches. Cette conception ignore le fait qu'un traitement sémantique minimal est très probable durant la répétition de mots ; les activités cérébrales correspondant à ce traitement minimal pourraient cependant être suffisantes pour diminuer le contraste entre les deux tâches, les différences régionales d'activation effectivement observées ne reflétant que très partiellement l'ensemble des régions cérébrales impliquées dans les processus sémantiques. (Cogn1 : composante cognitive de niveau 1 (ici transposition audio-phonatoire) ; Cogn 2. : composante cognitive de niveau 2 (association sémantique))

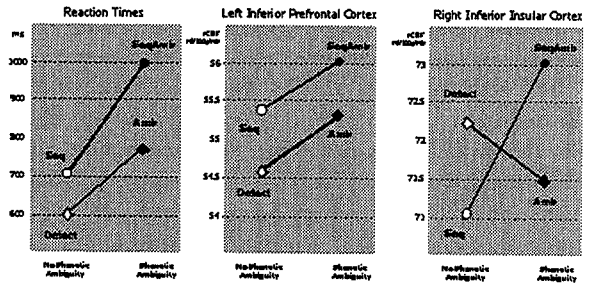


Figure 3. Non-linéarité de la dynamique d'activation et effets d'interaction entre facteurs psycholinguistiques. Plusieurs variantes de tâches de détection de phonèmes dans des pseudo-mots ont été étudiées dans un groupe de sujets. La variante la plus complexe (SeqAmb) consiste dans le repérage du phonème /b/si et seulement si le phonème /d/ a été perçu dans une syllabe précédente (exemple:/redozabu/) ; cette tâche combine deux facteurs de complexité : un traitement séquentiel des différentes syllabes des pseudo-mots et l'existence d'une ambiguïté phonétique entre phonèmes-cibles (/d/ et /b/) et distracteurs (/t/ et /p/). Une variante (Seq) ne comporte que le facteur séquentiel (détecter /b/ si /d/ avant , pas d'ambiguïté). Une autre variante (Amb) ne comporte que le facteur d'ambiguïté (pas de consigne relative à la prise en compte d'un phonème précédant le phonème cible /b/). Enfin, une 4ème variante (Detect) ne comporte aucun de ces facteurs de complexité. L'activation de l'aire de Broca répond à un modèle additive et à une dynamique d'activation strictement linéaire si l'on considère le taux d'activité enregistré dans cette région dans les 4 variantes : la combinaison des deux facteurs se traduit par une activation double de celle produite par chacun d'eux. Au contraire, la région inférieure du cortex insulaire droit est le lieu d'une interaction entre ces deux facteurs de telle sorte que le débit sanguin cérébral (rCBF) diminue par rapport à Detect lorsque l'on introduit l'un ou l'autre des facteurs (Amb ou Seq) alors qu'il augmente lorsqu'ils sont combinés (SeqAmb). Ainsi, l'insertion, du point de vue psychologique, de facteurs cognitifs dans une expérience ne se reflète pas toujours par l'addition de leurs effets respectifs du point de vue de leurs corrélats fonctionnels.

Avec la multiplication des travaux d'activation cognitive en PET, notamment ceux de l'équipe de R. Frackowiak à Londres, la constatation de résultats divergents, dans le domaine des activations liées au langage en particulier, a fait remettre en question cette conception quelque peu simpliste, bien que séduisante par la clarté des effets qu'elle permet de prévoir. En effet, d'une part des objections peuvent être formulées vis à vis de chacun des présupposés de la conception additive et hiérarchisée de l'activation et d'autre part, certains facteurs expérimentaux, initialement négligés, peuvent en fait influencer profondément les résultats.

La conception additive suppose possible la manipulation expérimentale de composantes cognitives, comme autant de "modules" présentant une indépendance les uns par rapport aux autres, tant du point de vue de leurs représentations psychologiques que du point de vue de leurs effets d'activation cérébrale. L'équipe londonienne a accumulé plusieurs exemples d'expériences (cf. notamment, Friston et al., 1996) dans lesquelles sont mis en évidence des **effets d'interaction entre les facteurs cognitifs** mis en œuvre, en contradiction avec les

prédictions de la conception additive de l'activation. Cette interaction signifie que, lorsque deux facteurs sont combinés dans l'une des tâches utilisées par l'expérience, **l'activation observée n'est pas égale à l'effet additionné de chacun de ces facteurs mais peut être supérieur à cette somme ou de direction contraire aux effets isolés de chaque facteur**. Ce type de résultat illustre une propriété fondamentale des phénomènes d'activation cérébrale liés aux fonctions cognitives: leur **dynamique non-linéaire**. Ces effets d'interaction sont en général observés dans certaines régions cérébrales, alors que d'autres, dans la même expérience, peuvent avoir au contraire une fonction de réponse conforme à une additivité des facteurs cognitifs (Figure 3). Ainsi, les fonctions de réponse du cortex vis à vis de tel ou tel facteur expérimental peuvent s'avérer très variables (passage d'une dynamique linéaire à une dynamique non-linéaire) pour des localisations corticales parfois très proches les unes des autres.

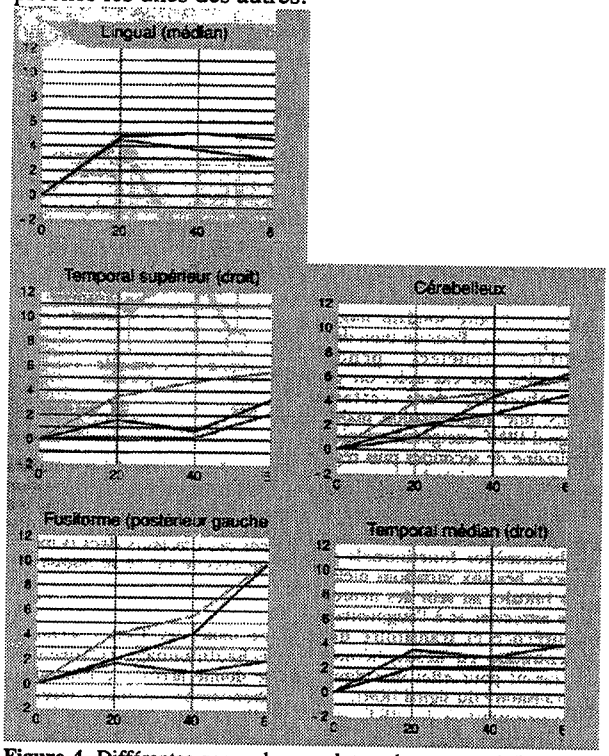


Figure 4. Différentes zones du cervelet et du cortex cérébral (régions associatives visuelles, occipitales et temporales) diffèrent radicalement par leurs fonctions de réponse à deux paramètres expérimentaux dans des tâches de lecture : 1) le rythme de présentation des stimulus (en abscisse, de 20 à 60 mots / min) et 2) leur durée de présentation sur l'écran (en bistre, présentation pendant 150 ms, lecture silencieuse ; en rouge, présentation de 1000 ms et lecture silencieuse ; en rose, présentation de 1000 ms et lecture à haute voix). En ordonnée : taux de variation du débit par rapport au repos. (d'après Price et al., 1996)

La complexité des phénomènes d'activation augmente encore si l'on considère maintenant l'influence des paramètres expérimentaux de l'activation cérébrale non plus seulement en termes qualitatifs (présence / absence de tel processus cognitif dans telle tâche) mais en termes quantitatifs.

Ainsi, dans une série de travaux liés notamment à l'activation cérébrale induite par des tâches de lecture, Cathy Price (de la même équipe londonienne) a montré que le fait de faire varier de manière systématique des paramètres expérimentaux tels que le nombre de stimuli présentés par minute (passant de 20 à 60 mots/min) ou la durée de présentation des stimuli (de 150 à 1000 ms), induit dans de nombreuses régions cérébrales (notamment dans les cortex associatifs occipitaux et temporaux) des variations **graduelles** d'activité, de telle sorte que l'activité dans une région corticale donnée soit considérée comme négligeable pour certaines valeurs de ces paramètres, alors qu'elle est majeure pour d'autres valeurs, transformant ainsi complètement les résultats de la comparaison tâche de lecture - tâche de référence. (Figure 4). De plus, ces résultats importants montrent qu'**il n'existe pas de relation univoque entre le sens de variation de tel paramètre psychologique et celui de l'activation corticale**: l'augmentation du rythme de présentation des stimuli peut, par exemple, faire augmenter le débit dans telle zone et le faire diminuer, ou le voir rester inchangé dans une zone voisine.

Ainsi, les **relations entre les caractéristiques cognitives des tâches et les effets d' »activation »** cérébrale peuvent donc être de type graduel. Par ailleurs, les comparaisons entre tâches d'activation doivent prendre l'influence **combinée** de **plusieurs** composants cognitifs non **pas indépendants** mais, au contraire, interagissant entre eux. Ces notions sont probablement de portée générale pour la physiologie de la cognition. Ainsi, au sein d'une étude d'activation, les relations entre différentes tâches cognitives et entre leurs corrélats neurofonctionnels, ne devraient plus être conçues simplement en termes d'opposition binaire pour un, et un seul, composant cognitif, relevant de son absence ou de sa présence. Au contraire, on doit plutôt considérer que **(a) chacune des tâches envisagées est en général susceptible de provoquer, en parallèle, la mise en jeu de tous les composants cognitifs pertinents et que (b) la charge attentionnelle ou le "degré d'engagement" relatifs à chacun de ces processus peuvent varier d'une tâche à l'autre**. Si l'on considère l'exemple d'une expérience linguistique comparant une tâche de repérage de phonèmes et une tâche de repérage de catégories sémantiques dans des mots, une conception strictement additive considérerait les processus phonologiques comme totalement enchâssés dans la tâche sémantique puisqu'un traitement phonologique est nécessairement inclus dans une tâche sémantique (il faut nécessairement « entendre » au moins partiellement les sons inclus dans un mot avant que de le comprendre). Cependant, le fait de demander au sujet de faire porter son attention soit sur le sens des mots soit sur leur structure interne en termes de contenu phonologique va faire en sorte que le "poids cognitif" ou la "charge attentionnelle" allouée aux processus phonologiques est moindre dans la tâche sémantique que dans la tâche phonologique et vice versa pour ce qui est des processus lexico-sémantiques. Ce gradient cognitif se traduit sur le plan de l'activation cérébrale par la présence de zones d'activation dans la tâche phonologique qui y induit alors une augmentation de débit plus importante

que ce qui est observé dans la tâche sémantique (Figure 5), contredisant ainsi la prédiction de la conception additive.

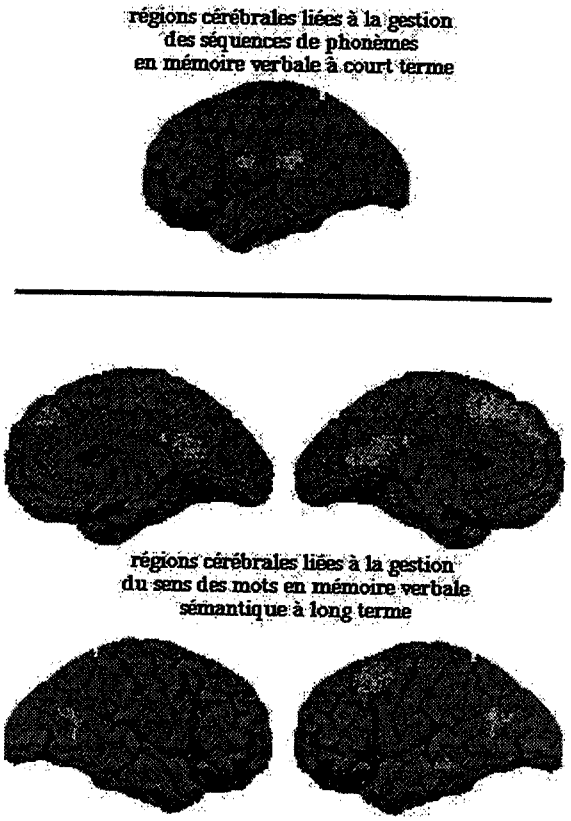


Figure 5. Débit sanguin cérébral enregistré lors d'une tâche de repérage de phonèmes. Vues transparentes de l'espace cérébral (Atlas de Talairach et Tournoux ; 1988), de profil droit (sagittal), d'arrière (coronal) et d'en haut (transverse, la gauche de la figure étant la gauche du cerveau). En bas à droite : vue standard externe de l'hémisphère cérébral gauche. Une augmentation significative de débit est induite par une tâche auditive de repérage de phonèmes (tâche SeqAmb cf. figure 3) dans le gyrus supra-marginal et le cortex moteur inférieur gauche, lorsque l'on considère comme tâche de référence une tâche auditive de repérage sémantique dans des mots. Ce surcroît d'activation dans la tâche de repérage de phonèmes par rapport à la tâche sémantique survient alors que des processus phonémiques existent obligatoirement dans cette tâche sémantique (la compréhension auditive des mots implique un minimum d'accès à leur structure phonémique). La différence ainsi mise en évidence entre ces tâches concerne donc le degré d'engagement de certains processus dans une tâche par rapport à une autre, plutôt qu'une opposition binaire en termes de présence/absence. (VPC : verticale de la commissure postérieure ; VAC : verticale de la commissure antérieure ; R : hémisphère cérébral droit) (d'après Démonet et al., 1994b)

Cette conception graduée des relations tâches/activations peut être exploitée pour l'analyse des résultats d'imagerie. Ainsi, dans une étude consacrée à l'apprentissage de listes de longueur progressivement croissante (passant de 2 à 12 mots), mettant donc en jeu la mémoire à court terme pour les listes infra-empan (environ 5 mots) et la mémoire à long terme pour les listes supra-empan (plus de 7 mots), Grasby et al. (1994) ont utilisé une analyse en composantes principales pour décrire les variations de débit induites par ces listes de longueur croissante. Sans faire intervenir aucun *a priori* (contrairement à la méthode soustractive) quant à la façon d'organiser les comparaisons entre les différentes mesures de débit effectuées , cette analyse de la variance observée dans l'ensemble des pixels du volume cérébral permet de mettre en évidence que, au fur et à mesure de l'allongement des listes à mémoriser, le profil d'activation cérébrale se modifie de telle sorte que l'activité bascule du système lié à la mémoire à court terme , de localisation péri-sylvienne, vers le système lié à la mémoire à long terme de localisation temporale médiane (Figure 6).

En résumé, l'observation de **fonctions de réponse corticale complexes** sous l'influence de tâches se distinguant entre elles par de multiples dimensions cognitives, illustre bien le caractère encore incomplet du modèle - en devenir - des relations cerveau/esprit tel que les résultats actuels d'imagerie fonctionnelle permettent de le concevoir.

Une partie des progrès à attendre dans l'élaboration de ce modèle proviendra sans doute de l'apport des deux techniques complémentaires de la PET dans ce domaine: l'IRM fonctionnelle et les techniques électro - ou magnéto-encéphalographiques.

Les signaux obtenus en IRM fonctionnelle (IRMf) grâce à la méthode d'imagerie fonctionnelle la plus répandue, la méthode BOLD (Blood Oxygenation Level Dependent, basée sur des variations de susceptibilité magnétique entre Oxy - et Déoxy - Hémoglobine), ont en fait des caractéristiques proches de celles de la PET dans la mesure où ils reflètent surtout des phénomènes d'ordre vasculaire, tels que des variations de débit sanguin local. L'avantage décisif de l'IRM fonctionnelle sur la PET est la possibilité d'effectuer des mesures d'activité en des temps extrêmement brefs (de l'ordre de 100 ms ou moins) et surtout de pouvoir répéter ces mesures à des intervalles courts, par exemple une mesure toutes les 3 ou 4 secondes. Cet échantillonnage relativement rapide donne accès à la dynamique du phénomène vasculaire essentiel lié à l'activation fonctionnelle: les modifications locales de débit et de volume de l'arbre micro - vasculaire entourant les synapses dont le métabolisme se trouve soudain modifié par l'expérience. Pour un train de stimulations répétées pendant 30 secondes par exemple, on peut ainsi observer en certaines régions, une augmentation majeure du signal IRM f atteignant un pic en une dizaine de secondes puis régressant, sur une période à peu près équivalente, vers des valeurs inférieures ou égales à celles observées avant toute stimulation. Un échantillonnage encore plus fin pourrait également permettre la mise en évidence de phénomènes fonctionnels plus précoces que ces effets vasculaires, liés à des variations métaboliques très précoces au sein des neurones et astrocytes, et à l'augmentation très précoce et transitoire de la concentration en Déoxy- Hémoglobine (ayant pour effet une diminution transitoire du signal IRM).

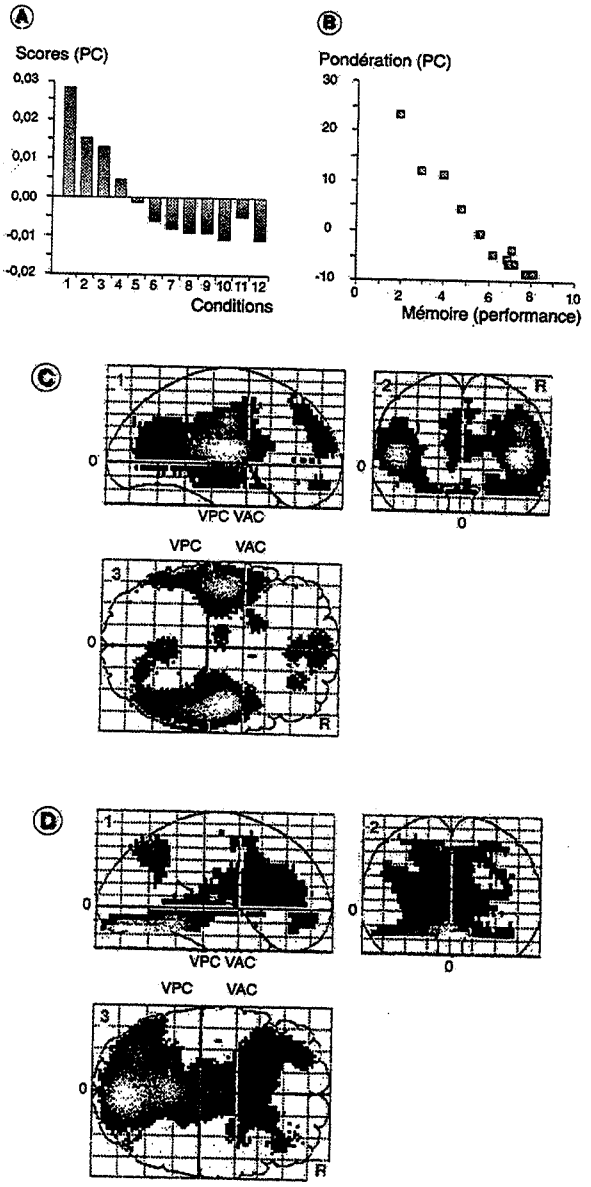


Figure 6. Anatomie fonctionnelle des systèmes de mémoire à court terme et à long terme explorée par Grasby et al. (1994) grâce à l'analyse en composantes principales de 12 mesures du débit sanguin cérébral en PET effectuées chez des sujets mémorisant des listes de mots de longueur progressivement croissante selon le numéro d'ordre des examens (de 1 à 12). Alors qu'elle ne suppose *a priori* aucune différence entre les mesures de débit, cette analyse révèle un premier vecteur (expliquant une part considérable de la variance totale) qui permet de discriminer les listes en fonction de leur longueur (et en fonction du nombre de mots rappelés par les sujets) et qui sépare en régions distinctes les dizaines de milliers de pixels examinés dans le volume cérébral. Le graphique en bas à gauche de la figure montre l' »intensité » et la « direction » de ce vecteur de variance dans chacune des 12 mesures PET. Ce vecteur est très positif pour les listes les plus courtes puis il passe par zéro pour les listes d'une longueur équivalente à l'empan (environ 5), pour ensuite prendre des valeurs de plus en plus négatives pour les listes les plus longues. En bas à droite de la figure est représentée la très forte corrélation négative entre les valeurs de ce vecteur et les performances moyennes de rappel des mots des listes observées chez les sujets au cours de cette expérience. En haut de la figure sont représentées, sous forme de cartes statistiques du cerveau dans l'espace de Talairach, les distributions dans le cerveau des pixels dont la variance est liée aux valeurs de ce vecteur. Les cartes « a » correspondent à une vue transparente de profil droit du cerveau, les cartes « b » sont des vues coronales et les cartes « c » sont des vues axiales. Les trois cartes en haut à gauche correspondent d'une part aux pixels co-variant pour les valeurs positives du vecteur et d'autre part aux listes courtes (mémoire à court terme). Les trois cartes en haut à droite correspondent à la direction négative de ce vecteur ainsi qu'aux listes longues (mémoire à long terme). Les pixels co-variants dans la direction positive et liés à la mémoire à court terme sont situés dans les régions péri-sylviennes (essentiellement temporales supéro-externes). Ceux associés à la direction négative du vecteur correspondent à la mémoire à long terme et aux régions temporales internes et diencéphalo-limbiques. Ces localisations cérébrales, ici identifiées selon une méthode statistique purement descriptive et "neutre" quant aux hypothèses relatives aux effets des tâches utilisées, correspondent d'assez près aux localisations des systèmes de la mémoire à court terme et de la mémoire à long terme établies à la lumière des travaux chez les patients cérébro-lésés.

Contrairement à la PET pour laquelle les comparaisons inter - tâches utilisent **une** valeur de débit mesurée par tâche, considérée comme un bloc, l'important échantillonnage des mesures IRM f au cours du temps expérimental permet d'analyser, pour chaque pixel des images obtenues, l'évolution temporelle du signal et donc de rechercher l'existence de corrélations entre ces variations de signal et l'évolution dans le temps des conditions expérimentales, c'est à dire, en général, l'alternance de condition "de repos" et de stimulations. La séquence de présentation de stimuli expérimentaux (par exemple, des trains de 30s comportant 10 stimuli alternant avec 30 s de "repos") peut alors être considérée comme une fonction d'entrée avec laquelle on recherche les pixels qui co-varient significativement, en prenant en compte un décalage de phase approprié puisque l'on sait que la vaso-réactivité mesurée possède une inertie de quelques secondes. Des travaux récents montrent la possibilité d'établir une relation encore plus étroite entre l'application de stimuli et l'observation de cette réactivité du lit micro - vasculaire à proximité des synapses fonctionnellement sollicitées. La technique décrite consiste à procéder à un moyennage de plusieurs mesures de débit rapidement réalisées à la suite de l'application de stimuli, ce qui revient à traiter le signal IRM f comme le signal électrique dans une étude de potentiels évoqués. Buckner et al. (1996a) et McCarthy et al. (1997) ont pu ainsi montrer l'intérêt de l'analyse de la dynamique spatio-temporelle des variations fonctionnelles du signal IRM grâce à cette technique d' »événements évoqués » (Figures 7a-d).

La mise au point de logiciels adaptés (SPM de Friston et Frackowiak), la multiplication des sites équipés d'imageurs IRM permettant les études fonctionnelles, ainsi que son caractère non-irradiant (autorisant, contrairement à la PET, la répétition d'études chez le même sujet) font désormais de l'IRMf la technique de choix pour les études dédiées à la "physiologie de la cognition" et ce en dépit des contraintes relativement lourdes que fait peser l'environnement magnétique sur la nature des équipements servant à réaliser les expériences d'activation.

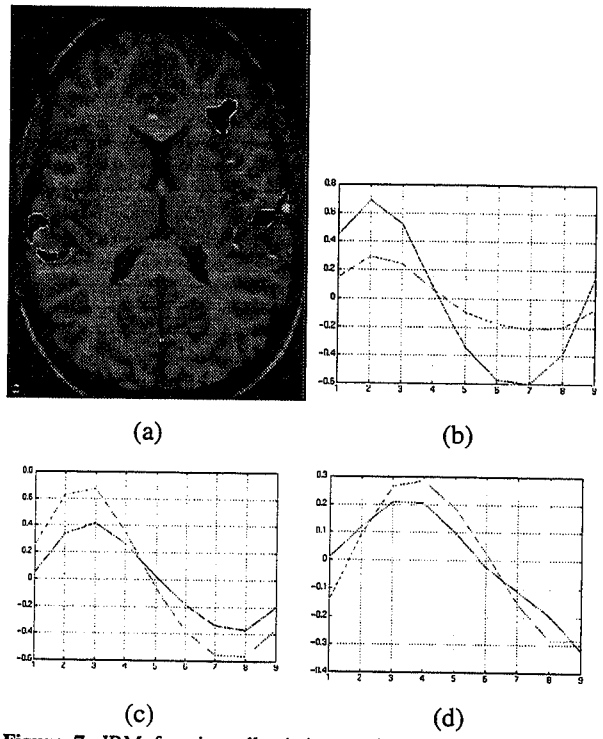


Figure 7. IRM fonctionnelle événementielle au cours de tâches de langage. (Thierry et al., 1999)

Figure 7a Carte statistique d'activation (p< .01) en IRMf événementielle chez un sujet (dont on visualise l'anatomie structurale) au cours d'une tâche phonologique de perception auditive et de catégorisation de suites de syllabes sans signification (pseudo-mots). Cette coupe axiale passe notamment par le cortex temporal supérieur et la région de l'opercule frontal (coordonnée z = 8mm selon l'atlas de Talairach et Tournoux). Trois groupes de pixels activés correspondent respectivement au cortex auditif associatif droit (gauche de l'image) et gauche et à l'aire de Broca.

Figure 7b Evénement hémodynamique évoqué dans la région du cortex auditif associatif (cortex temporal supérieur) droit par l'audition et le traitement phonologique de stimuli auditifs constitués de suites de syllabes sans signification (pseudo-mots) chez un sujet. Les courbes rouge et verte correspondent aux valeurs observées pour le pixel d'amplitude de variation maximum (p<.01) dans deux « runs » successifs ayant compris chacun 14 stimulations successives générant des signaux qui furent ensuite moyennés. Chaque numéro porté en abscisse correspond à une acquisition séparée de la suivante de deux secondes, en commençant une seconde après la délivrance du stimulus. Une échelle arbitraire des valeurs du signal IRMf normalisé sont portées en ordonnée.

Figure 7c. Evénement hémodynamique évoqué par la même tâche dans la région du cortex auditif associatif (cortex temporal supérieur) gauche (proche de l'aire de Wernicke). Mêmes conventions qu'en 7b.

Figure 7d. Evénement hémodynamique évoqué par la même tâche dans la région du cortex operculaire frontal gauche (proche de l'aire de Broca). Mêmes conventions qu'en 7b.

La comparaison des courbes en 7b, 7c et 7d montre que les pics d'activation dans l'aire de Wernicke (7c) et dans l'aire de Broca (7d) sont décalés respectivement d'environ deux et quatre secondes par rapport à celui observé dans le cortex auditif droit, témoignant peut-être d'un traitement plus prolongé de l'information linguistique dans ces deux aires dont on connaît l'implication pour le décodage et la programmation phonologique. Cette dernière, liée aux activités neuronales dans l'aire de Broca, s'accompagne de la réponse hémodynamique la plus tardive et la plus soutenue, ce qui pourrait traduire un phénomène de maintien de l'information verbale en mémoire de travail. Au contraire, l'activation induite par les stimuli dans le cortex auditif droit semble fugace comme si un traitement prolongé et spécifique ne pouvait s'y maintenir (voir à ce sujet Belin et al., 1998).

Grâce aux techniques d'Electro-Encéphalo-Graphie (EEG) et de Magnéto-Encéphalo-Graphie (MEG), combinées à la méthode des potentiels évoqués, des liens précis ont depuis longtemps été établis, dans le domaine temporel, entre variations des champs mesurés et des événements de type neurophysiologique comme l'intégration de stimulations sensorielles au niveau cortical, ou de type cognitif comme, par exemple, la détection d'événements rares parmi des stimuli fréquents, ou la détection d'incohérences sémantiques par rapport à un contexte. Ces deux derniers types d'expérience génèrent des événements neurophysiologiques (P300, N400) bien caractérisés et généralement considérés comme des corrélats neurophysiologiques d'étapes de traitement de l'information liées par exemple à l'attention sélective ou au traitement sémantique. Suivant une démarche inverse de celle des deux autres techniques (PET et IRM f), l'objectif technique dans le domaine de l'imagerie cérébrale électromagnétique est de mieux définir dans l'espace cérébral les sites ou générateurs responsables du déclenchement de ces variations de champ qui sont captées à distance.

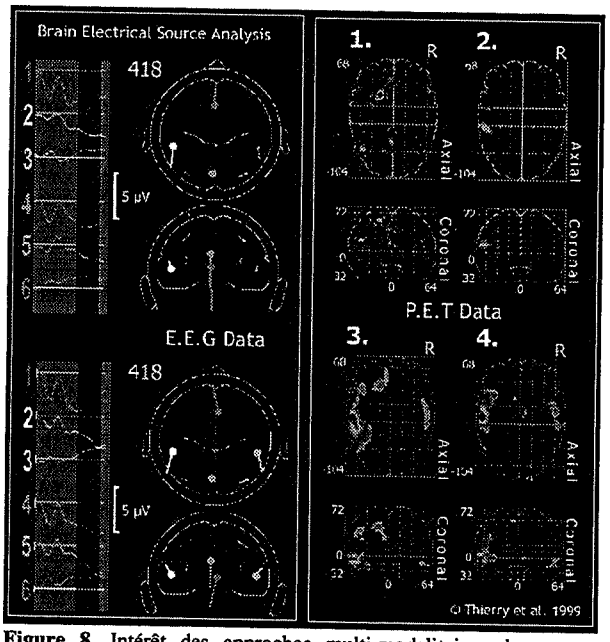


Figure 8. Intérêt des approches multi-modalitaires des processus cognitifs.

Une expérience en PET (Démonet et al., 1992 and 1994b) a eu pour but démontrer les corrélats des processus phonologiques et lexico-sémantiques dans la compréhension du langage oral. Six foyers d'activation distincts ont été identifiés dans les régions péri-sylviennes, temporales postérieures et cingulaires antérieures et postérieures. Ces deux derniers sites d'activation correspondent probablement à des composantes attentionnelles incluses dans les tâches. En ce qui concerne les composantes proprement linguistiques, elles se reflètent dans la prépondérance du foyer péri-sylvien gauche dans la tâche phonologique (cartes 1 et 2) alors que dans la tâche sémantique, on observe des foyers additionnels dans les régions temporales droites (cartes 3 et 4). La localisation de ces foyers a été utilisée pour modéliser les sources électriques des potentiels de surface enregistrés au cours d'une expérience similaire menée à l'aide des mêmes tâches linguistiques et de la technique de cartographie des potentiels évoqués

multi-canaux. La modélisation a été réalisée à l'aide du logiciel B.E.S.A. (Scherg, 1990). Outre la localisation des sources, le modèle incluait aussi une contrainte sur le fait que les sources homologues dans chaque hémisphère devraient avoir une orientation symétrique. Ce modèle permettait de rendre compte des données électrophysiologiques avec une variance résiduelle de seulement 0.35 % et ce durant une période de plus de 200 ms, dans une fenêtre de temps correspondant au déroulement des principales phases de traitement des informations phonologiques et sémantiques (pour des détails cf., Thierry et al., 1998).

Les sources de champ électrique ou magnétique sont constituées de macro-colonnes de cortex cérébral (3mm d'épaisseur et 3 mm de diamètre environ) au sein desquelles de nombreux axones sont arrangés de manière parallèle et sont excitées de façon synchrone, créant un champ dipolaire. Quelles que soient leurs orientations par rapport à la surface du scalp, l 'EEG capte les courants extra - cellulaires (de l'ordre de quelques micro - Volts au niveau du scalp) générés par ces structures. L'inconvénient essentiel du signal électrique est sa sensibilité à la conductance (variable) des milieux biologiques traversés, induisant déformation et diffusion du signal. La MEG capte des champs magnétiques générés par les courants intra - cellulaires. A la surface du scalp, ces champs sont de l'ordre de la centaine de femto-Tesla (1 million de fois plus faible que le champ magnétique terrestre). Par rapport au dipôle électrique généré dans une macro - colonne, le champ magnétique est orienté perpendiculairement. Les capteurs MEG disposés sur le scalp ne peuvent détecter que les macro - colonnes corticales orientées parallèlement à la surface du scalp, c'est à dire situées dans les sillons du cortex et non celles situées au "sommet" des gyri. En revanche, l'avantage de la MEG sur l 'EEG consiste dans le fait que le signal ne soit pas affecté par la traversée de milieux biologiques divers; le signal est donc peu déformé et diffusé. Il s'atténue avec la distance entre source et détecteur comme le carré de cette distance, donnée qui est exploitée dans le calcul de la localisation des sources.

Le fait d'avoir à identifier les sources intra - cérébrales de signaux recueillis à l'extérieur du crâne constitue la problématique essentielle de ce domaine. Il s'agit en fait de résoudre ce que l'on appelle en physique le "problème inverse" qui consiste en déterminer la localisation de la ou (plus souvent) des générateurs émettant les potentiels mesurés en surface. Ce problème n'a pas de solution mathématique unique; le nombre de variables susceptibles d'influencer les solutions est très important et l'estimation des solutions est extrêmement sensible à des erreurs minimes sur les données initiales.

Diverses méthodes sont utilisées pour circonscrire les solutions les plus plausibles. Une première façon est d'examiner le "problème direct" qui consiste dans le calcul des signaux recueillis en surface en considérant une source dont on connaît les caractéristiques et la localisation dans le cerveau. Il s'agit alors de modéliser la propagation de ce signal, à partir de leur source connue, dans les divers tissus céphaliques, en s'attachant particulièrement à l'anatomie cérébrale. La modélisation de la tête de chaque individu est réalisée à partir de l'IRM 3D et consiste en un maillage polyédrique très complexe, où chaque maille, unité volumique élémentaire du cerveau, du crâne et de la peau, fait l'objet d'un calcul de propagation.

On peut encore réduire le nombre de solutions en prenant en compte des données *a priori* . En se limitant à des phénomènes assez bien circonscrits tels que les potentiels liés à une modalité sensorielle particulière (auditive par exemple), on restreint la recherche à une région corticale spécifique. Une approche multi-modalitaire des expériences d'activation fonctionnelle trouve alors tout son intérêt (Figure 8). En effet, le fait de coupler dans les mêmes expériences le recueil de signaux MEG et celui de signaux EEG (comme le permettent les équipements multi-canaux les plus sophistiqués) permet d'affiner la localisation de sources. De plus, l'obtention préalable de données au moyen des autres modalités, tomographiques, d'imagerie fonctionnelle dans des expériences d'activation cognitive aussi comparables que possible, permet de postuler a priori la localisation des générateurs, en tout cas lorsque le nombre de foyers d'activation n'est pas trop grand et que les relations entre localisation des foyers et dynamique temporelle d'activation sont suffisamment claires (foyers proches des zones sensorielles ou motrices primaires). Dans les cas favorables, la résolution spatiale de la localisation de source devient alors très bonne, de l'ordre de quelques millimètres, et ce au niveau de l'anatomie cérébrale d'un individu donné .

Enfin, une approche complémentaire, permettant d'améliorer encore la localisation des sources et d'enrichir en général nos connaissances sur la physiologie cognitive, consiste en une approche des signaux MEG et EEG en termes de "cohérence" fréquentielle, afin de déterminer quelles sont les régions cérébrales qui présentent des variations corrélées - ou des co-variations - des signaux qu'elles émettent au cours du temps expérimental et en fonction des contraintes comportementales qu'il suppose. De nouveau, l'intérêt de cette approche co-variationnelle des données MEG et EEG réside dans le fait qu'elle peut également s'appliquer à l'imagerie tomographique PET et IRM f, comme l'ont souligné Friston et collaborateurs (1993). Cette approche commune aux différentes imageries pourrait alors être la base d'une véritable compréhension de la physiologie de la cognition en fusionnant son caractère distribué dans l'espace cérébral et sa labilité sur la flèche du temps.

En dépit du caractère encore très imparfait de notre compréhension des relations cerveau/esprit à travers les méthodes d'imagerie fonctionnelle (et leur éventuelle combinaison), il est évident que leur apport est déjà considérable. Le nouveau paradigme n'en évince pas pour autant l'ancien, celui des lésions. Les deux paradigmes sont au contraire complémentaires et des travaux récents ont d'ores et déjà montré l'intérêt de leur combinaison permettant d'étudier, par exemple, chez des aphasiques effectuant des tâches d'activation, la fonctionnalité des territoires cérébraux épargnés (par les lésions et/ou par

leurs effets à distance), ainsi que l'existence de relations entre l'activation de ces territoires sains et le maintien (ou la récupération) de fonctions linguistiques. Ce champ de recherche, actif, voit actuellement s'opposer deux thèses, l'une soutenant que l'hémisphère droit joue un rôle dans la récupération des fonctions linguistiques chez les aphasiques (Buckner et al., 1996b ; Weiller et al., 1995; Musso et al., 1999), l'autre mettant l'accent sur le rôle crucial des régions épargnées de l'hémisphère gauche (Belin et al., 1996 ; Warburton et al., 1999), spécialement la partie postérieure du lobe temporal gauche (Heiss et al., 1999). La méthode fonctionnelle/lésionnelle devrait ainsi permettre de mieux rendre compte des dissociations entre processus perturbés et épargnés par les lésions et de mieux comprendre les mécanismes cérébraux et cognitifs présidant à la vicariance des processus perturbés et à la récupération fonctionnelle en général.

POUR EN SAVOIR PLUS...

Frackowiak RSJ et Friston K (1995) Methodology of activation paradigms. In Handbook of Neuropsychology, Vol 10 (R. Johnson Jr, J-C Baron, Section Editors) (F. Boller, J. Grafman, series editors), Elsevier Science B.V, pp. 369-382.

Démonet J-F (1998) Tomographic brain imaging of language functions : prospects for a new brain/language model. In Stemmer B, Whitaker HA, eds. *Handbook.of Neurolinguistics*. San Diego : Academic Press, 1998 : 132-143.

Dehaene S. et collaborateurs, Le cerveau en action. Imagerie cérébrale fonctionnelle en psychologie cognitive. Collection "Psychologie et Sciences de la Pensée", *PUF*.

La Recherche, N° de Juillet - Août 1996 "Voir dans le cerveau".

REFERENCES

Belin P., Van Eeckhout Ph., Zilbovicius M., Rémy Ph., François C., Guillaume S., Chain F., Rancurel G., Samson Y. (1996). Recovery from nonfluent aphasia after melodic intonation therapy: a PET study. Neurology, 47:1504-1511.

Belin P, Zilbovicius M, Crozier S, Thivard L, Fontaine A, Masure MC, Samson Y. (1998) Lateralization of speech and auditory temporal processing. *J Cogn Neurosci*, 10: 536-40.

Buckner RL, Bandettini PA, O'Craven KM, Savoy RL, Petersen SE, Raichle ME, Rosen BR. (1996a) Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* ; 93 : 14878-14883.

Buckner RL, Corbetta M, Schatz J, Raichle ME, Petersen SE. (1996b) Preserved speech abilities and compensation following prefrontal damage. *Proc. Natl. Acad. Sci. USA* ; 93 : 1249-1253.

Caramazza A. (1996) Pictures, words and the brain. News and Views. *Nature*, 383: 216-217.

Démonet, J.F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.L., Wise, R., Rascol, A., and R.S.J. Frackowiak (1992) The anatomy of phonological and semantic processing in normal subjects, *Brain*, 115 : 1753-1768.

Démonet J-F, Price C, Wise R, Frackowiak RSJ (1994a) A PET study of cognitive strategies in normal subjects during language tasks: influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain*, 117; 671-682.

Démonet J-F, Price C., Wise R., Frackowiak RSJ. (1994b) Differential activation of right and left posterior sylvian regions by semantic and phonological tasks: a positron-emission tomography study in normal human subjects. Neuroscience Letters, 182:25-28.

Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ (1993) Functional connectivity: The principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow and Metabolism* :13 ; 5-14.

Friston K.J., Price C.J., Fletcher P., Moore C., Frackowiak RSJ., Dolan R.J. (1996) The trouble with cognitive subtraction. NeuroImage, 4: 97-104.

Grasby P., Frith C.D., Friston K.J., Simpson J., Fletcher P.C., Frackowiak R.S.J., Dolan R.J. (1994) A graded task approach to the functional mapping of brain areas implicated in auditory-verbal memory. Brain, 117: 1271-1282.

Heiss WD, Kessler J, Thiel A, Ghaemi M, Karbe H (1999) Differential capacity of left and right hemispheric areas for compensation of poststroke aphasia. *Ann Neurol* 45:430-8.

Jeannerod M. (1996) De la physiologie mentale. Histoire des relations entre la psychologie et la biologie. Eds Odile Jacob.

Jenkins I.H., Brooks D.J., Nixon P.D., Frackowiak RSJ., Passingham RE. (1994) Motor sequence learning: a study with positron emission tomography. The Journal of Neuroscience, 14: 3775-3790.

Lassen NA, Ingvar DH. (1961) The blood flow of the cerebral cortex determined by radioactive Krypton-85. Experientia., 17 : 42-43.

McCarthy G, Luby M, Gore J, Goldman-Rakic P. (1997) Infrequent events transiently activate human prefrontal and parietal cortex as measured by functional MRI. *J. Neurophysiol.*, 77 : 1630-1634.

Messerli P. (1993) Une approche historique de l'aphasie. In: Langage et aphasie (Eds: F. Eustache, B. Lechevalier), DeBoeck Université, pp. 13-39.

Musso M, Weiller C, Kiebel S, Muller SP, Bulau P, Rijntjes M (1999) Training-induced brain plasticity in aphasia. *Brain* 122 :1781-1790.

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* : 331; 585-589.

Price C.J., Moore C.J., Frackowiak R.S.J. (1996) The effect of varying stimulus rate and duration on brain activity during reading. NeuroImage, 3, 40-52.

Roy C, Sherrington C. (1890) On the regulation of the blood supply of the brain. *J Physiol*, 11 : 58-108.

Schulman G.L., Fiez J.A., Corbetta M., Buckner R.L., Miezin F.M., Raichle M.E. and Petersen S.E. (1997). Common Blood Flow Changes across Visual Tasks: II. Decreases in Cerebral Cortex. J Cog Neurosci. 9:648-663.

Talairach J, Tournoux P (1988) *Co-Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. Translated by Mark Rayport. New York: Thieme Medical Publishers, Inc. Stuttgart, New York: George Thieme Verlag.

Thierry G., Doyon B. and Démonet J.F. (1998) ERP Mapping in Phonological and Lexical Semantic Monitoring Tasks: a Study Complementing Previous PET Results. NeuroImage. Nov;8(4):391-408.

Thierry G., Boulanouar K., Kherif F., Ranjeva J-P. and Démonet J-F. (1999) Sorting neural components underlying phonological processing. Neuroreport 10: 2599-2603.

Vandenberghe R., Price C., Wise R., Josephs O., Frackowiak RSJ. (1996) Functional anatomy of a common semantic system for words and pictures. *Nature*, 383: 254-256.

Warburton E, Price CJ, Swinburn K, Wise RJ (1999) Mechanisms of recovery from aphasia: evidence from positron emission tomography studies. *J Neurol Neurosurg Psychiatry* 66:155-61.

Weiller C, Isensee C, Rijntjes M, Huber W, Müller S, Bier D., Krams M, Faiss JH, , Noth J, Diener HC (1995) Recovery from aphasia after stroke. A positron emission tomography study. *Annals of Neurology* : 37, 723-732.

Indexation, Segmentation et Analyse de Scènes

Analyse en composantes principales temps-fréquence : application à la reconnaissance de la langue

Michel Dutat[(1)(2)], Ivan Magrin-Chagnolleau[(3)], Frédéric Bimbot[(3)]

[(1)] LSCP / CNRS UMR 8554, 54 Boulevard Raspail, 75270 Paris cedex
[(2)] ENST - Dépt Signal, CNRS URA 820, 46 Rue Barrault, 75634 Paris cedex 13
[(3)] IRISA / CNRS & INRIA Rennes, Campus universitaire de Beaulieu, 35042 Rennes cedex
dutat@lscp.ehess.fr - ivan@ieee.org - bimbot@irisa.fr

ABSTRACT

In this paper, we use a new speech parameterization based on a principal component analysis applied to feature parameters augmented by their context. This new parameterization is called time-frequency principal component (TFPC) analysis. We apply the new parameterization in the framework of automatic language recognition. This new approach allows us to improve the identification rate compared to the use of the classical cepstral coefficients augmented by their Δ coefficients.

1. INTRODUCTION

Une grande variété d'analyses paramétriques du signal de parole a été utilisée en reconnaissance de la langue par modélisation acoustique. Les meilleurs résultats sont généralement obtenus en incorporant l'information dynamique du signal par le biais d'approximations de la dérivée et de la dérivée seconde (les coefficients Δ et $\Delta\Delta$). Dans cette étude, nous abordons une nouvelle paramétrisation qui prend également en compte l'aspect dynamique du signal. Cependant, cette méthode opére sur le signal une sélection à la fois temporelle et fréquentielle du matériel acoustique et ceci en fonction de la langue. Cette sélection est réalisée par des filtres temps-fréquences dont les coefficients sont calculés à partir du corpus d'apprentissage de la langue considérée. On fait l'hypothèse qu'un énoncé ainsi filtré est mieux représenté lorsque le filtre qui lui est appliqué est celui correspondant à sa langue. Cette approche rend plus optimal la paramétrisation et améliore ainsi le taux de reconnaissance par rapport à celui obtenu avec une paramétrisation classique.

2. PRÉSENTATION DE LA MÉTHODE

2.1. *Approche classique*

Lors d'une procédure de reconnaissance de la langue par une modélisation acoustique, on tente de capter la structure sonore propre à cette langue en utilisant un grand nombre d'enregistrements prenant en compte le plus de locuteurs possible. Vient ensuite la paramétrisation des enregistrements dans une forme acceptable pour les modèles. Ceux-ci sont en général constitués de réseaux de Markov cachés à structure ergodique. La figure 1 présente la procédure classique d'une phase d'apprentissage. L'ensemble des énoncés d'apprentissage $E_{app}^{(i)}$ de la langue i est transformé par une analyse acoustique en une séquence de vecteurs de paramètres $\{\mathbf{x}_t\}^{(i)}$. Cette séquence est utilisée afin d'estimer les paramètres $\kappa^{(i)}$ du modèle de la langue i.

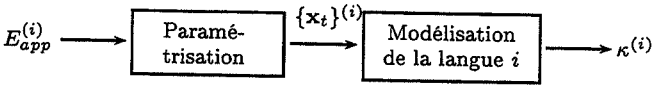


Figure 1 – *Apprentissage classique.*

Une fois que tous les paramètres des modèles ont été calculés, on obtient n modèles, chacun représentant une langue différente. La phase de test consiste à paramétriser l'énoncé à tester, puis à calculer une mesure de confiance par les n modèles du système. Un algorithme de décision combine alors les n scores obtenus, avec éventuellement d'autres sources d'informations, afin de sélectionner la langue la plus probable.

2.2. *Approche par filtrage vectoriel*

La méthode que nous présentons a été initialement proposée dans [4, 5]. Elle diffère de l'approche classique à la fois en ce qui concerne l'apprentissage des modèles et en ce qui concerne la procédure de test. Son but est de rendre l'énoncé plus facilement identifiable dès l'étape de paramétrisation. Les énoncés, après une analyse acoustique, subissent un filtrage vectoriel temps-fréquence dont les caractéristiques sont dépendantes de la langue. Ainsi un énoncé filtré avec le filtre de sa langue est mieux représenté que s'il l'est par un tout autre filtre. On obtient ainsi une paramétrisation qui doit être plus optimale. La phase d'apprentissage comporte deux étapes. La première permet l'obtention des coefficients du filtre et la seconde permet le calcul des paramètres $\lambda^{(i)}$ du nouveau modèle. La figure 2 illustre l'apprentissage du modèle de la langue i. L'ensemble $E_{app}^{(i)}$ des énoncés d'apprentissage est transformé par une analyse acoustique[1] en une séquence $\{\mathbf{x}_t\}^{(i)}$ de vecteurs de dimension p. Celle-ci permet le calcul des coefficients de la matrice de filtrage, puis une fois les caractéristiques du filtre $\mathbf{H}^{(i)}$ obtenues, le filtrage de la séquence $\{\mathbf{x}_t\}^{(i)}$ en une nouvelle séquence $\{\mathbf{f}_t\}^{(i)}$ de vecteurs de dimension r permet le calcul des paramètres $\lambda^{(i)}$ du modèle de la langue i.

1. qui peut être spectrale, cepstrale, par prédiction linéaire, etc.

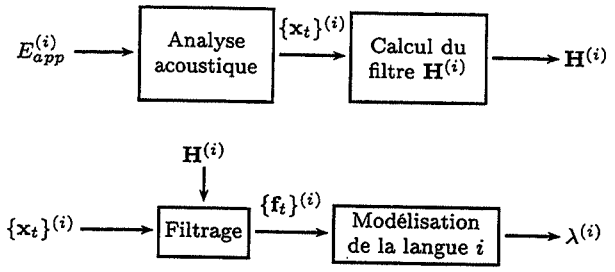


Figure 2 – *Apprentissage avec filtrage vectoriel.*

2.3. *Fabrication du filtre*

Les caractéristiques du filtre sont capitales pour obtenir des vecteurs bien répartis dans l'espace de représentation des langues. Ainsi, pour une langue donnée, et à partir de l'analyse acoustique de ses données d'apprentissage, on recherche une représentation des données qui maximise l'inertie, dans un sous espace qui sera caractéristique de la langue. Pour ce faire, nous utilisons une analyse en composantes principales de la séquence des vecteurs d'apprentissage. Le but est de rechercher les liaisons qui existent entre les différents coefficients des vecteurs paramètres et qui peuvent caractériser la langue en question. De plus, pour capter l'information dynamique résidant dans la suite de vecteurs consécutifs, on va également prendre en compte un contexte temporel autour de chaque trame analysée. On obtient en fin de compte un filtrage vectoriel à base de composantes principales temps-fréquences (que l'on nommera par le sigle TFPC *Time Frequency Principal Components*)[4, 5].

Les différentes étapes pour l'élaboration d'un filtre sont les suivantes : soit la séquence de vecteurs $\{\mathbf{x}_t\}$, de dimension p, issue d'une analyse acoustique des énoncés d'apprentissage, avec t variant de 1 à T, on construit les matrices de covariances décalées V_q avec $q = 0, 1, 2, \cdots$:

$$\mathsf{V}_q = \frac{1}{T}\sum_{t=q+1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}})(\mathbf{x}_{t-q} - \overline{\mathbf{x}})^{\mathsf{T}} \quad \text{avec} \quad \overline{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$$

Puis on construit la matrice de covariance contextuelle $\mathbf{V}_{2q+1}$ à partir des matrices de covariances décalées de telle sorte que l'on obtienne une matrice de structure bloc-Toeplitz de dimension $(2q+1)p \times (2q+1)p$. On appellera ordre de l'analyse le nombre q qui prend en compte $2q+1$ trames temporelles.

$$\mathbf{V}_{2q+1} = \begin{bmatrix} \mathsf{V}_0 & \mathsf{V}_1 & \cdots & \mathsf{V}_{2q} \\ \mathsf{V}_1^{\mathsf{T}} & \mathsf{V}_0 & \cdots & \mathsf{V}_{2q-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{V}_{2q}^{\mathsf{T}} & \mathsf{V}_{2q-1}^{\mathsf{T}} & \cdots & \mathsf{V}_0 \end{bmatrix}$$

L'analyse en composantes principales des données augmentées de leur contexte temporel revient à rechercher les vecteurs propres $\mathbf{v}_i$ de la matrice $\mathbf{V}_{2q+1}$. L'espace de représentation composé des vecteurs propres correspondant aux plus grandes valeurs propres est celui d'inertie maximale. Il est à noter que toutes les directions de cet espace sont orthogonales entre elles. De plus, les composantes principales de moindres variances peuvent être assimilés à du bruit. Il est ainsi possible de diminuer l'espace de représentation en ne conservant que certains vecteurs propres représentant aux mieux les données. Il existe d'ailleurs plusieurs stratégies pour réduire la dimension de l'espace de représentation [3]. À partir des composantes principales retenues, on élabore une matrice de filtrage en juxtaposant ces vecteurs et en la transposant. Ainsi, si l'on veut les cinq premières composantes suivies de la 10e à la 12e, on construit la matrice de filtrage comme suit :

$$\mathbf{H} = [\mathbf{v}_1 \cdots \mathbf{v}_5\ \mathbf{v}_{10} \cdots \mathbf{v}_{12}]^{\mathsf{T}}$$

Le filtrage d'un énoncé consiste en un produit de convolution entre sa représentation paramétrique et la matrice $\mathbf{H}$. Soit la séquence de vecteurs $\{\mathbf{x}_t\}_{1\leq t\leq T}$ de taille p issue de l'analyse acoustique de l'énoncé. On définit la séquence de vecteurs centrés $\mathbf{x}_t^*$ entre le temps $t-q$ et le temps $t+q$:

$$\mathbf{X}_{t-q}^{t+q} = \begin{bmatrix} \mathbf{x}_{t+q}^* \\ \vdots \\ \mathbf{x}_t^* \\ \vdots \\ \mathbf{x}_{t-q}^* \end{bmatrix} \qquad \text{avec } \mathbf{x}_t^* = (\mathbf{x}_t - \overline{\mathbf{x}})$$

La dimension du vecteur $\mathbf{X}_{t-q}^{t+q}$ est $(2q+1)p \times 1$. En supposant que nous ayons choisi r composantes principales, la matrice de filtrage, de dimension $r \times (2q+1)p$, peut s'écrire en faisant apparaître sa structure temporelle :

$$\mathbf{H} = [\mathsf{H}_{-q} \cdots \mathsf{H}_0 \cdots \mathsf{H}_q]$$

Les vecteurs filtrés $\mathbf{f}_t$ sont obtenus par le produit de convolution suivant avec $1 \leq t \leq T$:

$$\begin{aligned} \mathbf{f}_t &= \mathbf{H} \cdot \mathbf{X}_{t-q}^{t+q} \\ &= \sum_{k=-q}^{+q} \mathsf{H}_k \cdot \mathbf{x}_{t-k}^* \end{aligned}$$

2.4. *La phase de test*

Une fois la phase d'apprentissage terminée, à chaque langue est associée un modèle mais également un filtre. Tester l'origine linguistique d'un énoncé revient à faire calculer par chaque modèle une mesure de vraisemblance. Mais ici la paramétrisation de l'énoncé subit, pour chaque modèle, le filtrage correspondant. La figure 3 illustre la procédure de test employée avec la méthode TFPC. Une analyse acoustique traduit l'énoncé de test E_{test} de langue inconnue en une séquence de vecteurs $\{\mathbf{x}_t\}$. Pour chacune des n langues du système, cette séquence est filtrée. Puis une estimation de la probabilité d'observation conditionnelle $P\left(\{\mathbf{f}_t\}|\lambda^{(i)}\right)$ est calculée. L'étape de décision revient à sélectionner la langue $L^{(i)}$ de plus forte probabilité.

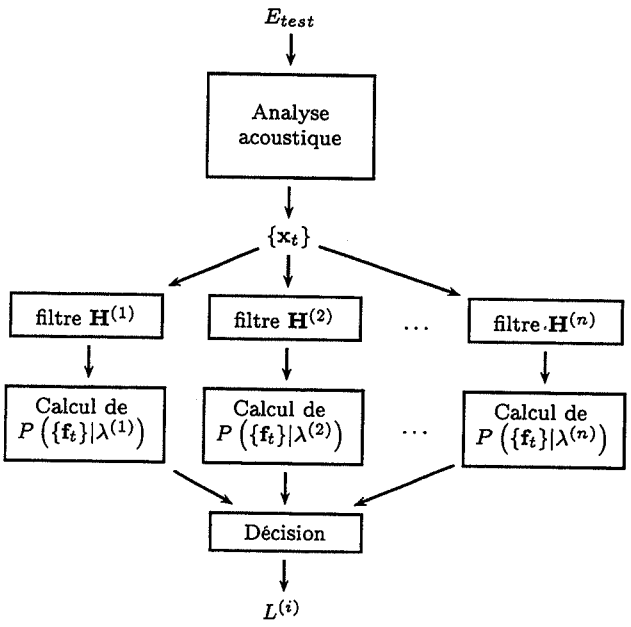


Figure 3 – *Procédure de test en utilisant la méthode* TFPC.

Cette méthode offre un large éventail de possibilités quant à la paramétrisation finale des énoncés. Ces choix portent sur les points suivants :

- Analyse acoustique des énoncés. (bancs de filtres, coefficients lpc, coefficients cepstraux, etc.)
- Choix de l'ordre de l'analyse TFPC, c'est-à-dire le nombre de trames temporelles prises en compte.
- Choix des vecteurs de la matrice des composantes principales à faire intervenir pour l'élaboration du filtre.

Dans les sections suivantes, on présente des résultats obtenus en faisant varier certains de ces paramètres. Une expérience classique faisant intervenir une paramétrisation cepstrale des énoncés nous permet d'évaluer l'efficacité de cette nouvelle approche.

3. Résultats

3.1. Objectif

Nous désirons étudier l'influence de cette nouvelle paramétrisation sur les taux de succès d'un système de reconnaissance de la langue. Pour ce faire, on construit deux systèmes de reconnaissance qui ne diffèrent que par leur étape de paramétrisation des énoncés, l'un appelé modèle classique, l'autre appelé modèle TFPC. Nos systèmes doivent attribuer à l'énoncé testé une langue parmi les langues apprises.

3.2. Corpus

Nous avons constitué nos corpus d'apprentissage et de test en choisissant des enregistrements dans la base de données OGI MLTS (Oregon Graduate Institute Multilanguage Telephone Speech) qui a été spécialement conçue pour des travaux de reconnaissance de la langue [6]. Pour cette étude, nous avons choisi quatre langues : l'anglais, l'espagnol, le français et le japonais. Pour chaque langue, nous avons utilisé pour l'apprentissage (resp. pour le test) 80 énoncés (resp. 54) prononcés par 20 locuteurs (resp. 12), ce qui représente une durée totale de 1240 secondes (resp. 860).

3.3. Caractéristiques du modèle classique

Ce modèle de contrôle nous permet d'analyser l'apport de cette nouvelle paramétrisation. Il consiste en une modélisation de la structure acoustique de la langue à partir d'un apprentissage non supervisé. Il est constitué d'un modèle de Markov caché par langue dont les caractéristiques sont données dans la table 1.

Table 1 – *Caractéristiques du modèle de Markov caché:*

Structure	ergodique
Nombre d'états	24
Nombre de gaussiennes par état	2
Matrices de covariances	diagonales

Pour le choix de la paramétrisation du signal de parole, nous avons écarté l'analyse par bancs de filtres, des études antérieures ayant montré que les résultats d'un système de reconnaissance utilisant cette paramétrisation sont moins bons que ceux utilisant une analyse cepstrale [1]. On a donc choisi cette dernière pour la mise en forme des énoncés du modèle de contrôle. Les caractéristiques de cette analyse sont résumées dans la table ci-dessous :

Table 2 – *Caractéristiques de l'analyse cepstrale :*

Type de la paramétrisation	MFCC [a]
Répartition fréquentielle	échelle Mel
Nombre de coefficients	12 ou 24
Largeur de la fenêtre d'analyse	30 ms
Période d'analyse	10 ms
Fenêtre d'analyse	Hamming
Soustraction cepstrale	oui

[a] Mel Frequency Cepstral Coefficient

Comme certaines expériences avec le modèle TFPC font intervenir le contexte temporel dans la paramétrisation, il faut le faire également intervenir dans le modèle classique. Ceci est réalisé par l'apport des coefficients différentiels Δc_t^i qui sont calculés à partir des coefficients cepstraux c_t^i par la formule [2] :

$$\Delta c_t^i = \frac{\sum_{k=1}^{N} k(c_{t+k}^i - c_{t-k}^i)}{2 \cdot \sum_{k=1}^{N} k^2} \quad \text{avec} \quad i = 1, 2 \cdots, p$$

La variable N définit l'horizon temporel à prendre en compte pour le calcul des coefficients différentiels. Suivant l'ordre de l'expérience TFPC, on aura l'équivalence suivante :

Ordre	N	Nombre de c^i	Nombre de Δc^i
0	-	24	0
1	1	12	12
2	2	12	12
3	3	12	12
4	4	12	12

3.4. Expériences TFPC

Pour comparer les résultats, il faut que la différence entre les deux systèmes ne porte que sur le filtrage des

énoncés. Ainsi, la configuration du modèle de Markov caché pour le modèle TFPC est identique à celle du modèle de contrôle (cf. Table 1). Le modèle classique utilisant une paramétrisation cepstrale, nous gardons le même type d'analyse acoustique pour les expériences TFPC (cf. Table 2). Afin d'analyser l'apport de l'information dynamique sur les taux de succès de reconnaissance, on fait varier l'ordre de l'analyse TFPC en prenant soin de modifier en conséquence la paramétrisation de l'expérience classique de telle sorte que l'horizon temporel pris en compte soit identique pour les deux expériences. On s'intéresse également au nombre de composantes principales retenues pour l'élaboration des filtres TFPC. Ainsi, selon les expériences, on sélectionne de 16 à 32 composantes prises parmi les premiers vecteurs propres de la matrice $\mathbf{V}_{2q+1}$.

Ordre 0 L'analyse TFPC d'ordre 0 n'utilise pas de contexte temporel. La matrice de covariance contextuelle $\mathbf{V}_{2q+1}$ est donc la matrice de covariance V_0 de dimension 24×24. La matrice de filtrage sera au maximum de même dimension et les vecteurs filtrés auront donc aux maximum 24 coefficients. Par conséquent, l'expérience classique doit également avoir un vecteur de paramètres composé de 24 coefficients. Le tableau ci-dessous présente les différents taux de succès moyen sur les quatre langues considérées obtenus avec l'expérience TFPC en fonction du nombre de composantes retenues pour fabriquer le filtre. La dernière ligne est composée des gains relatifs par rapport à l'approche classique qui dans cette configuration a obtenu un taux de succès moyen de 53,10%. On constate une amélioration des résultats et celle-ci d'autant plus forte que l'on diminue le nombre de composantes du filtre. Il semble que l'effet du filtrage TFPC ait bien permis de retenir les informations dépendantes de la langue, mais si les premières composantes principales sont efficaces, les dernières semblent avoir des caractéristiques partagées par plusieurs langues, ce qui expliquerait l'accroissement des confusions lorsque l'on augmente la taille de l'espace de représentation.

Composantes	16	20	24
Taux de Succès	58,02	55,78	53,47
Gain relatif	9,27	5,05	0,70

Ordre 1 Avec un ordre $q = 1$, le contexte temporel s'étend sur trois trames du signal. La matrice de covariance contextuelle $\mathbf{V}_3$ est de dimension 72×72. L'expérience de contrôle fait intervenir l'estimation de la dérivée première dans sa paramétrisation du signal, avec un horizon temporel de 3 trames ($N = 1$). Son taux de succès moyen sur quatre langues est de 59,94%. Le tableau ci-dessous résume les résultats obtenus. Comme précédemment, une amélioration est à mettre au compte de la méthode TFPC, mais l'effet dû à la diminution de l'espace de représentation est moins marqué. Il semble que l'information pertinente soit plus confusément répartie à travers les composantes principales et une étude plus poussée du choix de celles-ci semble nécessaire.

Composantes	16	20	24	28	32
Taux de Succès	62,71	60,81	57,67	59,99	61,65
Gain relatif	4,62	1,45	-3,79	0,08	2,85

Ordre 4 L'ordre $q = 4$ fait intervenir 9 trames temporelles. La matrice de covariance contextuelle $\mathbf{V}_9$ a pour dimension 216×216. La paramétrisation de l'expérience classique comporte les coefficients ΔMFCC calculés avec $N = 4$. Le taux de succès moyen pour l'expérience de contrôle est de 60,34%. Le tableau des résultats confirme l'effet de confusion constaté précédemment. L'information utile pour la discrimination des langues ne peut être exploitée en n'utilisant qu'une trentaine de composantes. Cependant avec 32, mais surtout 24 composantes retenues, on obtient une amélioration par rapport à l'expérience de contrôle.

Composantes	16	20	24	28	32
Taux de Succès	53,91	58,55	63,49	53,59	60,75
Gain relatif	-10,66	-2,97	5,22	-11,19	0,68

4. Conclusions

Nous avons abordé une nouvelle paramétrisation qui, à partir d'une analyse acoustique classique, applique sur les vecteurs de paramètres un filtrage vectoriel dépendant de la langue. Les caractéristiques des filtres sont obtenues par une analyse en composantes principales des données d'apprentissage. Par ce procédé, on utilise des espaces de représentations optimisés pour chaque langue. Les expériences en reconnaissance automatique de la langue que nous avons réalisées montrent une légère augmentation des taux de succès lorsqu'on applique cette méthode. L'un des avantages de cette approche est la grande liberté que nous avons quant aux choix concernant la paramétrisation initiale, la taille du contexte temporel à prendre en compte, la sélection des vecteurs de la matrice des composantes principales pour l'élaboration du filtre. Cependant les expériences impliquant un contexte temporel important nous ont montré la nécessité d'un choix plus subtil des composantes à intégrer aux matrices de filtrage. On s'intéressera également à l'influence d'une telle paramétrisation sur la régularité des séquences d'états parcourus dans les modèles de Markov cachés. Ceci est particulièrement important dans l'optique de travaux sur la localisation de suites de séquences d'états typiques propres aux langues.

Références

[1] Michel Dutat. Reconnaissance automatique de la langue parlée. Rapport d'avancement de thèse, Ecole National Supérieure des Télécommunications, 1997.

[2] Sadaoki Furui. Comparison of speaker recognition methods using static features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):342–350, June 1981.

[3] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[4] Ivan Magrin-Chagnolleau. *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte*. PhD thesis, Ecole National Supérieure des Télécommunications, 1997.

[5] Ivan Magrin-Chagnolleau, Geoffrey Durou, and Frédéric Bimbot. Application of time-frequency principal component analysis to text-independent speaker identification. Submitted to IEEE Transactions on Speech and Audio Processing.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. Technical report, Center for Spoken Language Understanding Oregon Graduate Institute of Science and Technology., Portland, 1993.

Utilisation des moments d'ordre 3 pour une détection Parole/non-Parole robuste

Arnaud Martin

France Télécom R&D/DIH/DIPS
2, avenue Pierre Marzin – 22307 Lannion, France
Tél.: ++33 (0)296 05 23 10 - Fax: ++33 (0)296 05 35 30
Mél: arnaud.martin@rd.francetelecom.fr

ABSTRACT

In noisy environment, robustness to noise of speech/non-speech detection is necessary for speech recognition. This paper presents a new method for speech/non-speech detection using third-order statistics. The analysis of the energy third-order statistic behavior gives useful information on the energy distribution. The new algorithm is evaluated in terms of segmentation and recognition performance. Different telephone call environments are considered for the evaluation. This algorithm is compared to the one based on noise and speech statistics presented in [Kar98a]. The results show that the new algorithm outperforms the one based on second order noise and speech statistics only, especially in the case of noisy environment.

1. INTRODUCTION

Les performances de la reconnaissance vocale décroissent en utilisation dans des environnements très bruités. Une détection efficace des périodes de parole et de non-parole est cruciale pour la reconnaissance.

C'est dans ce sens que de nombreuses études ont été menées. La caractéristique principale du signal utilisée est l'énergie, mais on peut lui associer par exemple la fréquence fondamentale [Iwa99] pour détection plus fine. Un algorithme de détection de parole/non-parole fondé sur l'utilisation de l'estimation des statistiques de l'énergie des périodes parole et des périodes de non-parole, a été présenté dans [Kar98]. Cet algorithme nous servira d'algorithme initial. Des méthodes comme la logique floue peuvent également être utilisées [Ber99].

L'observation que la distribution du signal de parole est non gaussienne a conduit différentes études à considérer les statistiques d'ordre supérieur dans des systèmes de détection de la parole. Dans [Jac91], il est proposé l'utilisation du skewness et du kurtosis (qui sont les cumulants d'ordre 3 et 4 normalisés). [Dou97] utilise le fait que le cumulant croisé de deux variables aléatoires indépendantes soit nul, pour discriminer le signal parole et celui du bruit à la source. Dans [Nem99], les auteurs intègrent le cumulant d'ordre 4 du résidu des coefficients de prédiction LPC dans un système de détection d'activité vocale. On se propose ici, d'intégrer le moment d'ordre 3, normalisé, conditionnellement à l'algorithme initial utilisant les statistiques de l'énergie dans des périodes de parole et de non-parole. Cette statistique permet une description plus fine de la distribution de l'énergie qui fournit la décision dans le système de détection Parole/non-Parole.

Ce papier est organisé comme suit. Dans la section 2, on rappellera l'algorithme initial de détection parole/non-parole, utilisant les statistiques de l'énergie des périodes de parole et de non-parole. Dans la section 3, est présentée l'estimation des moments d'ordre supérieur, et en particulier du moment d'ordre 3. On introduit ensuite un nouveau critère de détection de parole et de non-parole fondé sur le moment d'ordre 3, que l'on intègrera dans l'algorithme initial. Dans la section 4, on présente les performances de cet algorithme de détection en comparaison avec l'algorithme initial. Les évaluations sont faites sur deux bases de données téléphoniques, enregistrées l'une à travers le réseau RTC, l'autre à travers le réseau GSM, dans différents environnements d'appel.

2. ALGORITHME INITIAL

On rappelle dans cette section l'algorithme de détection de parole/non-parole, fondé sur un automate à cinq états [Mau94]. Les cinq états de l'automate sont : *silence, présomption de parole, parole, plosive ou silence et reprise possible de la parole.* Les transitions d'un état à l'autre se font à l'aide de contraintes de durée, et par des tests sur les caractéristiques du signal. Ces différents tests correspondent à différents critères.

Les transitions d'un état à l'autre de l'automate se font par un test d'hypothèse sur chaque trame du signal observé. On considère la moyenne et la variance de l'énergie dans les périodes de parole et dans celles de non-parole [Kar98]. L'estimation des statistiques de l'énergie des périodes de parole se fait dans l'état *parole*, celles des périodes de non-parole dans l'état *silence*. Cette approche découle d'une approche Bayésienne. On teste deux hypothèses :

H_0 : on est dans un état de non-parole,

H_1 : on est dans un état de parole.

On considère dans un premier temps que les distributions de l'énergie dans les périodes de parole et dans celles de non-parole suivent deux lois normales.

La décision du passage d'un état à l'autre de l'automate se fera, pour chaque trame observée x par comparaison du maximum de vraisemblance $P(H_i / x)$ de chaque hypothèse, pour i=0 ou 1. En supposant les deux hypothèses équiprobables, et utilisant la formule de Bayes, on se ramène à la comparaison à un seuil du rapport de vraisemblance :

$$R(x) = \frac{P(x/H_0)}{P(x/H_1)}.$$

3. CRITERE DU MOMENT D'ORDRE 3

L'étude expérimentale du rapport du moment d'ordre 3 dans les périodes de parole et de bruit montre que c'est un paramètre pertinent pour affiner la description de la distribution de l'énergie du bruit et de la parole. La décision faite uniquement sur la moyenne et l'écart type de l'énergie va donc pouvoir être précisée.

3.1 Estimation des statistiques d'ordre supérieur

L'estimation "classique" des moments d'ordre r, $\hat{\mu}_r$ de l'énergie est l'estimation arithmétique :

$$\hat{\mu}_r(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^r$$

où x_i est l'énergie du signal pour la $i^{ème}$ trame, et N est le nombre de trames. Cette estimation a l'inconvénient de ne pas tenir compte de la non-stationnarité du signal de parole. On utilise donc ici une estimation sur des fenêtres exponentielles, qui revient à pondérer les trames avec des poids décroissant avec le temps. Ainsi pour une trame donnée n, le moment d'ordre r $\hat{\mu}_r$ est définit par :

$$\hat{\mu}_r(n) = \hat{\mu}_r(n-1) + (1-\lambda)(x_n^r - \hat{\mu}_r(n-1))$$

où λ est le facteur d'oubli. Le degré supposé de stationnarité du signal détermine le facteur λ, et ainsi le nombre de trames qui seront considérées pour le calcul de la statistique. Cet estimateur contrairement à l'estimateur arithmétique n'est pas sans biais, en effet on a :

$$E[\hat{\mu}_r(n)] = (1-\lambda^{n+1})\mu_r$$

où μ est le moment d'ordre r théorique. Cet estimateur est asymptotiquement sans biais. L'étude de la variance de ces estimateurs reste un problème délicat quant à l'établissement des formules générales. Une étude de ces estimateurs pour les grandes valeurs de n a été réalisée dans [McC87]. Il est montré que ces estimateurs sont consistants, avec une vitesse de convergence diminuant avec la croissance de l'ordre du moment.

3.2 Le moment d'ordre 3

On considère ici le moment d'ordre 3, normalisé, définit par :

$$\hat{m}_3(n) = \frac{\hat{\mu}_3(n)}{\hat{\sigma}^3(n)}$$

où $\hat{\mu}_3$ et $\hat{\sigma}$ sont respectivement l'estimateur du moment d'ordre 3 et de l'écart type de l'énergie. Ces estimateurs sont calculés de la façon décrite dans la section 3.1, sur des fenêtres exponentielles. Le moment d'ordre 3, normalisé, d'une quantité de moyenne nulle est exactement le skewness. De plus, la variance de cet estimateur reste suffisamment faible à l'échelle de l'hypothèse de stationnarité du signal, elle est notamment plus faible que celle du skewness et celle du moment d'ordre 3 centré.

3.3 Intégration du moment d'ordre 3

Nous allons voir comment intégrer ce paramètre dans l'algorithme initial. $\hat{m}_3$ est calculé avec deux facteurs d'oubli différents : l'un ($\lambda_{cl} = 0.9$) qui donnera une estimation à court terme du moment d'ordre 3, l'autre ($\lambda_{ll} = 0.99$) qui donnera une estimation à long terme. L'estimation à long terme est calculée récursivement uniquement dans les périodes de non-parole, c'est-à-dire dans l'état *silence* de l'automate. La décision sera prise en comparant le rapport $rap(n) = \frac{\hat{m}_{3cl}(n)}{\hat{m}_{3ll}(n)}$, des estimations à court terme et à long terme du moment d'ordre 3, à un seuil adaptatif. Ce rapport est plus faible dans les périodes de parole. Cette décision est prise conditionnellement au test sur l'énergie de l'algorithme initial, et aux contraintes de temps. C'est-à-dire que pour le passage d'un état à l'autre, on comparera d'abord le rapport $R(x)$ à un seuil, la décision sera confirmée par le test du moment d'ordre 3. Le seuil adaptatif est calculé récursivement dans les périodes détectées comme de la parole par l'algorithme initial, de la façon suivante :

$$\hat{T}(n) = \hat{T}(n-1) + (1-\lambda_T)(c.rap(n-1) - \hat{T}(n-1))$$

où λ_T est un facteur d'oubli, et $c > 1$ est un coefficient permettant d'obtenir une borne supérieure du rapport des moments dans les périodes de parole. Il a été optimisé à une valeur proche de 3 sur nos bases de données

Cette méthode suppose deux hypothèses, les distributions énergétiques du bruit et de la parole ont des moments d'ordre 3 différents, et le bruit est plus stationnaire que la parole.

4. EXPERIMENTATIONS

Les tests ont été effectués sur deux bases de données. Pour évaluer le nouvel algorithme on a procédé à une évaluation de la segmentation et à une évaluation de la reconnaissance à partir d'un système de reconnaissance élaboré au CNET [Mok97].

4.1 Bases de données

Une première base de données est constituée de 1000 appels téléphoniques à un serveur vocal interactif en exploitation, donnant les programmes de cinéma. Les appels enregistrés en continuité à travers le réseau RTC contiennent les mots de commande au serveur (soit un vocabulaire de 25 mots). Le corpus obtenu par la segmentation manuelle contient 67% de mots du vocabulaire, 11% de parole hors vocabulaire et 22% de bruit.

La deuxième base de données est une base enregistrée par téléphone (hors contexte applicatif), constituée de 51 mots de vocabulaire que chaque locuteur répète. Les 395 appels ont été effectués à travers le réseau GSM, à partir de différents environnements (*intérieur, extérieur, voiture à l'arrêt, voiture roulant*). Le corpus a été segmenté manuellement, 68% des segments sont des mots du vocabulaire, 4% des mots hors vocabulaire et 28% des bruits.

4.2 Test de segmentation

Les tests de segmentation sont effectués par comparaison à la base segmentée manuellement. Les segments de parole du vocabulaire, hors vocabulaire et différents types de bruits ont été annotés. Ainsi différentes erreurs apparaissent, les omissions les insertions, les regroupements et les fragmentations [Mau94]. En vue de la reconnaissance ces erreurs sont classées en erreurs rejetables (comportant les insertions et les détections de parole hors vocabulaire) et en erreurs définitives (comportant les omissions, les fragmentations et les regroupements). Les courbes sont obtenues en faisant varier le seuil de détection de l'algorithme initial (seuils indiqués sur les courbes).

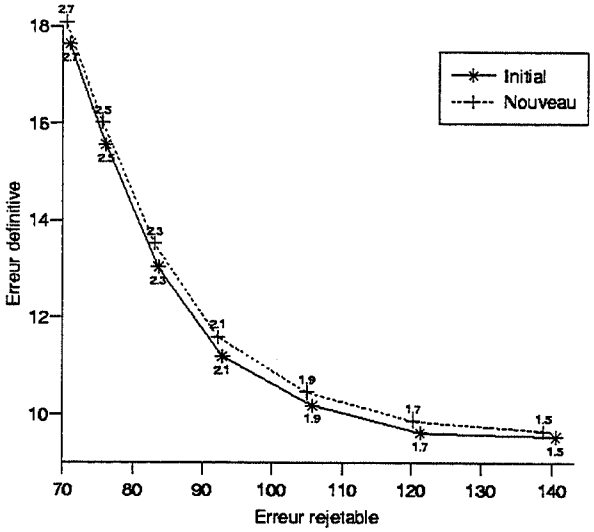


Figure1: Test de segmentation – données RTC

La figure 1 présente les résultats des tests sur la base de données enregistrée à travers le réseau RTC. Dans cet environnement non bruité, l'algorithme initial présente des points de fonctionnement conduisant à moins d'erreurs pour la reconnaissance. On remarque que pour un même seuil le nouvel algorithme donne davantage d'erreurs définitives (dû aux fragmentations), mais un peu moins d'erreurs rejetables (dû aux insertions).

La figure 2 donne les résultats des tests sur la base de données enregistrées à travers le réseau GSM. On a regroupé ici les quatre environnements. Le nouvel algorithme présente un peu moins d'erreurs en vue de la reconnaissance. Lorsque les tests sont effectués séparément sur chaque environnement l'écart est plus prononcé pour les environnements les plus bruités (*extérieur, véhicule roulant*).

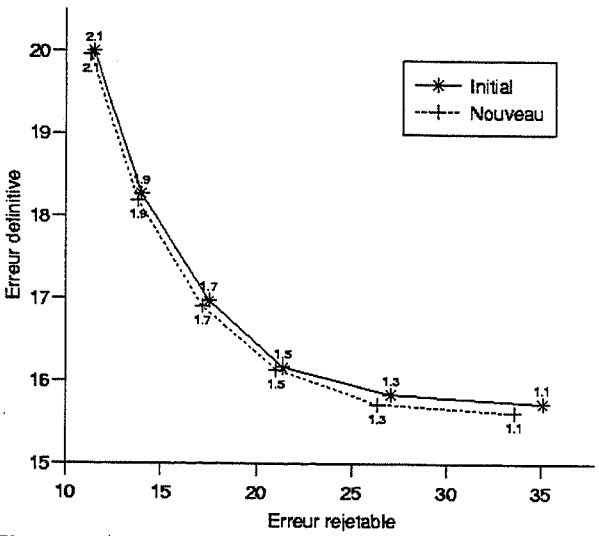


Figure2: Test de segmentation – données GSM

4.3 Test de reconnaissance

Le système de reconnaissance utilisé au CNET [Mok97] est fondé sur la modélisation des mots du vocabulaire à partir des chaînes de Markov. Pour modéliser toutes les réalisations acoustiques possibles, on utilise un modèle par allophones (modélisation contextuelle des phonèmes). Cette modélisation est faite à partir d'une base différente de celles utilisées pour les tests. Les courbes des tests de reconnaissance sont obtenues en faisant varier l'importance du rejet. On représente ici les erreurs de substitution associées aux fausses acceptations en fonction des faux rejets. Pour chaque algorithme deux seuils ont été choisis pour la détection.

Les résultats des tests sur la base enregistrée à travers le réseau RTC sont équivalents pour les deux algorithmes. Le peu d'erreurs définitives en plus est compensé par moins d'erreurs rejetables.

La figure 3 donne les résultats des tests sur la base de données enregistrée à travers le réseau GSM. Le nouvel algorithme présente des taux d'erreur légèrement plus faibles que l'algorithme initial. La différence se situe surtout au niveau des fausses acceptations. Ceci vient du fait qu'il y a moins d'erreurs rejetables. La figure 4 montre que ces résultats sont encore plus accentués

dans le cas d'un environnement bruité où les enregistrements ont été effectués lors dans un véhicule roulant.

4.4 Discussion

Les résultats obtenus montrent une légère amélioration du nouvel algorithme dans des conditions bruitées. Ceci vient du fait que le nouveau critère a été utilisé conditionnellement au critère de l'algorithme initial. Cependant cette amélioration reste peu significative. En effet l'écart des taux d'erreur reste très faible. La diminution des erreurs se retrouve surtout au niveau des erreurs rejetables. Les omissions et les fragmentations ne sont pas diminuées (les erreurs définitives). Cette approche diminue cependant les fragmentations dans des environnements très bruités.

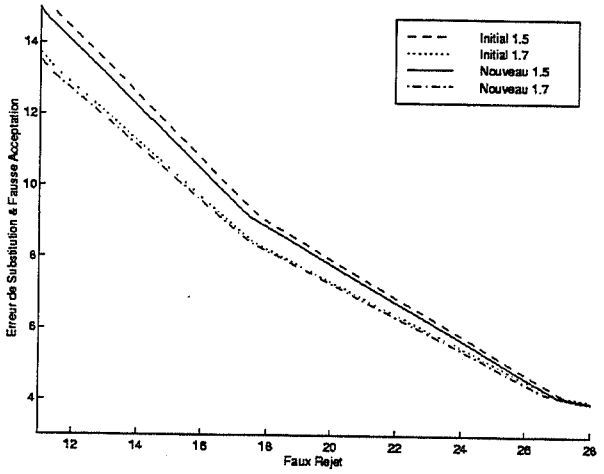


Figure3: Test de reconnaissance – données GSM

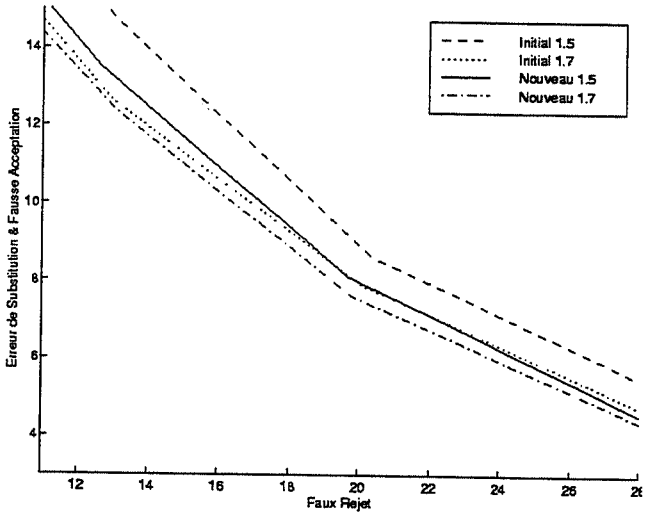


Figure4: Test de reconnaissance – données GSM véhicule roulant

5. CONCLUSION ET PERSPECTIVES

Nous avons intégré un critère sur les moments d'ordre 3 de l'énergie du signal dans l'algorithme de détection de parole/non-parole utilisant les statistiques des périodes de parole et de non-parole. Le rapport des moments d'ordre 3 à court terme et à long terme nous a permis de comparer les distributions de l'énergie des périodes de parole et de non-parole. Les tests de segmentation et de reconnaissance ont montré une légère amélioration de l'algorithme utilisant ce nouveau critère. Cette amélioration reste cependant peu significative.

Nous avons ici introduit les moments d'ordre 3 conditionnellement à la décision prise par l'algorithme initial, dans l'état de *parole* de l'automate. Ce nouveau critère pourrait conduire à une autre approche. Nous nous proposons dans une prochaine étude de combiner une décision fournit par ce nouveau critère avec la décision prise par l'algorithme initial, par le biais des méthodes de fusion de décision.

BIBLIOGRAPHIE

[Ber99] Beritelli F., Casale S. and Cavallaro K. (1999), "A Multi-Channel Speech/Silence Detector based on Time Delay Estimation and Fuzzy Classification", ICASSP, Vol. 1, pp. 93- 96.

[Iwa99] Iwano K. and Hirose K. (1999), "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition", ICASSP, Vol. 1, pp. 133- 136.

[Jac91] Jacovitti G., Pierucci P. and Falashi A.(1991), "Speech Segmentation and Classification Using Higher Order Moments", Eurospeech, pp. 1371-1374.

[Dou97] Doukas N., Naylor P. and Stathaki T. (1997), "Voice Activity Dection Using Source Separation Techniques", Eurospeech, pp. 1099-1102.

[Kar98] Karray L. and Monné J. (1998) "Robust speech/non-speech detection in adverse conditions based on noise and speech statistics", ICSLP, Vol.. 4, pp. 1471-1474.

[Mau94] Mauuary L. (1994) "Amélioration des serveurs vocaux interactifs", Thèse de Doctorat, université de Rennes 1.

[McC87] McCullagh P. (1987) "Tensor Methods in Statistics", Charpman and Hall.

[Mok97] Mokbel C. *et al.* (1997) "Towards improving ASR robustness for PSN and GSM telephone applications", Speech communication, Vol. 23, pp 141-159.

[Nem99] Nemer E., Gourbran R. and Mahmoud S. (1999) "The fourth-order cumulant of speech signals with application to voice activity detection", Eurospeech, Vol. 5, pp. 2391-2394.

Séparation de sources de parole : une nouvelle approche utilisant la cohérence audiovisuelle des signaux

L. Girin, A. Allard, G. Feng, & J.-L. Schwartz

Institut de la Communication Parlée – INPG/Univ. Stendhal/CNRS
Domaine Universitaire – BP 25 - 38040 Grenoble Cedex 9, France
Tél.: ++33 (0)476 82 41 20 - Fax: ++33 (0)476 82 43 35
girin@icp.inpg.fr – http://www.icp.inpg.fr

ABSTRACT

We present a new approach to the source separation problem, in the case of a mixture of speech signals. The method is based on the use of the visual component of speech gestures : the goal is to extract the audio component thanks to its coherence with lip movements. An original algorithm is introduced, together with a statistical model of audiovisual coherence. Preliminary results are quite satisfactory: they show that it is indeed possible to separate an audiovisual speech source with this method, which presents some interesting complementarity with traditional pure audio techniques.

1. INTRODUCTION

Dans de nombreuses situations, la voix d'un interlocuteur peut être en partie couverte par des bruits environnants ou par les voix d'autres personnes (effet *"cocktail-party"*), ce qui en atténue notablement la compréhension. Dans ce genre de problèmes, on désire souvent isoler la voix d'un locuteur particulier voire même un à un l'ensemble des signaux. Ceci est typiquement un problème de séparation de sources, domaine du traitement du signal qui a récemment suscité de nombreuses études avec un certain nombre d'applications en parole [Com94].

Ces études utilisent toutes les seules propriétés audios des signaux de parole (généralement des propriétés d'indépendances entre signaux provenant de sources différentes) sans s'intéresser à la capacité des êtres humains à extraire de l'information en observant le visage de leur locuteur. En effet, la parole est un vecteur de communication à la fois auditif et visuel [Sto96], et l'homme est capable d'utiliser cette bimodalité pour "isoler" relativement bien la voix d'un interlocuteur, c'est-à-dire la faire ressortir perceptivement par rapport aux autres voix ou par rapport à un bruit environnant.

Une étude récente menée à l'ICP a permis de montrer la faisabilité d'un système automatique de rehaussement de la parole dans un bruit blanc utilisant l'image du locuteur [Gir98]. Le travail présenté dans ce papier est un premier prolongement de cette étude au cas plus complexe du mélange additif instantané de plusieurs sources de parole. Il s'agit d'une tentative d'élaboration d'un système de séparation de sources de parole utilisant automatiquement de l'information relative au(x) visage(s) parlant(s), plus précisément des paramètres géométriques décrivant la forme des lèvres du (des) locuteur(s). Le principe général de ce système est décrit à la section 2, et son évaluation sur un premier corpus de test est présentée à la section 3.

2. PRESENTATION DU SYSTEME

2.1. Principe théorique

En entrée du système, on a un ensemble de signaux observations x_i résultant du mélange additif de plusieurs signaux de parole acoustique a_j. Ces signaux sont bien sûr fonction du temps, mais on omet la variable t pour simplifier les notations. Par souci de simplicité, prenons le cas du mélange de deux signaux avec deux capteurs :

$$\begin{cases} x_1 = m_{11}a_1 + m_{12}a_2 \\ x_2 = m_{21}a_1 + m_{22}a_2 \end{cases}$$

Les m_{ij} sont des coefficients multiplicatifs inconnus supposés constants pendant le processus de séparation. Le principe général de nombreux systèmes de séparation de sources consiste à estimer chaque signal source par une combinaison linéaire des observations, soit

$$\begin{cases} s_1 = c_{11}x_1 - c_{12}x_2 \\ s_2 = c_{22}x_2 - c_{21}x_1 \end{cases}$$

puis à déterminer les coefficients de démêlage c_{ij} pour que les signaux s_j soient proches des a_j originaux. Notons que dans le cas N signaux et N capteurs, on ne doit déterminer que Nx(N-1) coefficients si on récupère les s_i à un facteur de gain près. Ainsi, pour le système 2x2, on cherche

$$\begin{cases} s_1 = x_1 - c_1x_2 \approx \alpha a_1 \\ s_2 = x_2 - c_2x_1 \approx \beta a_2 \end{cases}$$

La plupart des systèmes de séparation classiques cherchent la solution qui rend les s_j indépendants deux à deux puisque les a_j le sont eux-mêmes. Dans le système présenté ici, les observations x_i sont accompagnées d'un signal vidéo v_j contenant l'information vidéo relative au locuteur j et synchrone au signal a_j que l'on désire isoler. Ce signal contient la trajectoire de paramètres géométriques décrivant le contour labial du locuteur. Dans cette étude, il s'agit des deux paramètres basiques d'étirement (E) et de hauteur (H) du contour labial

interne. Ces paramètres peuvent être extraits automatiquement par divers systèmes développés au laboratoire avec une période de 20 ms. Le principe du système audiovisuel consiste à renverser le paradigme habituel des systèmes de séparation de sources classiques. Plutôt que de rechercher un critère d'indépendance entre sources acoustiques, on veut ici tirer profit de la cohérence intrinsèque entre les signaux de parole audio et vidéo issus d'un même locuteur. En d'autres termes, le critère de réglage des coefficients de démêlage c_j sera la maximisation de la cohérence – dans un sens que nous allons définir – entre le signal acoustique estimé s_j et la source vidéo v_j d'un même locuteur.

La particularité du problème que nous avons à résoudre est que l'information fournie par le canal vidéo est incomplète, et notamment qu'elle ne peut contribuer qu'à la spécification du *spectre* de la source, mais en aucun cas de ses variations temporelles fines. Il faut donc définir un algorithme capable d'exploiter cette information spectrale. Une étude préliminaire [All99] nous a permis d'élaborer un tel algorithme dans un cadre théorique simplifié, où l'on suppose que l'on connaît parfaitement le spectre de l'une des sources. Nous avons pu ainsi montrer que l'on pouvait obtenir un débruitage parfait et rapide sur des mélanges 2 signaux / 2 capteurs ou 4 signaux / 4 capteurs, et même dans certaines conditions sur des mélanges avec plus de sources que de capteurs : l'information spectrale permet de se focaliser sur le signal à débruiter, ce qui représente un atout intéressant par rapport aux techniques classiques de séparation de sources.

Dans la pratique, nous n'avons pas accès directement au spectre de la source, mais seulement à une estimation de ce spectre dans un cadre probabiliste. Il s'agit alors d'obtenir, pour chaque couple de signaux audio et vidéo (a, v) dans une phase d'apprentissage, une estimation de la probabilité audiovisuelle associée $p(A, v)$ où A est le spectre de a. Dans la phase de test, face à un couple (x, v) où x est bruité, on définit un signal débruité s de spectre S, et c'est la probabilité $p(S, v)$, que l'on cherche à maximiser. C'est cette modélisation statistique de la cohérence des signaux de parole audiovisuelle que nous allons décrire maintenant.

2.2. *Modélisation statistique de la cohérence audiovisuelle*

Le principe de cette modélisation est d'associer à des vecteurs audiovisuels (A, v) (A étant le spectre de a) une probabilité $p(A, v)$. Pour cette première étude, la modélisation statistique choisie est un modèle standard de mixture de gaussiennes. La probabilité audiovisuelle est ainsi donnée par une somme pondérée de N lois gaussiennes multi-dimensionnelles, soit

$$p(A, v) = \sum_{k=1}^{N} \frac{G_k}{\sqrt{(2\pi)^d \det C_k}} e^{-\frac{1}{2}([A\ v]-M_k)C_k^{-1}([A\ v]-M_k)'}$$

où d est la dimension de l'espace audiovisuel (voir section 3). M_k, C_k et G_k sont respectivement le vecteur moyenne, la matrice de covariance et le poids associés à la gaussienne k. L'algorithme utilisé pour déterminer ces paramètres est celui d'*Expectation-Maximisation* (EM). Son principe est de calculer ces caractéristiques de manière itérative à partir d'un corpus de données d'apprentissage. On fixe N au départ de l'algorithme et on donne des conditions initiales arbitraires pour les paramètres des gaussiennes. L'idée est ensuite d'alterner une phase où les données sont classifiées par recherche du maximum de la probabilité (Maximisation) et une phase où les paramètres des gaussiennes sont réajustés en fonction des résultats de ce classement (Expectation). Ce processus itératif est mené jusqu'à ce que le système de gaussiennes obtenu n'évolue plus. Cette phase d'apprentissage étant faite, on peut associer à tout nouveau couple (spectre A, paramètres labiaux v) de l'espace audiovisuel une probabilité $p(A, v)$.

3. EXPERIMENTATION

3.1. *Données*

Pour tester la faisabilité de cette technique nouvelle, nous nous sommes placés dans des conditions bien contrôlées, avec un corpus composé de séquences $V_1CV_2CV_1$. V_1 et V_2 sont des voyelles parmi [a, i, y, u], et C est une consonne plosive parmi [p, t, k, b, d, g]. Ces séquences sont prononcées par un même locuteur, avec deux répétitions. Il s'agit d'un corpus de référence déjà bien étudié à l'ICP du point de vue perceptif et utilisé dans le cadre du débruitage audiovisuel [Gir98]. Ce corpus monolocuteur constitue donc le seul signal à isoler a_1 pour notre système, tandis que le signal perturbant a_2 est constitué de phrases du français issues d'un corpus audio standard utilisé pour des tests en codage de parole.

Les sons sont échantillonnés à 16kHz. L'enveloppe des spectres est fournie par une analyse par prédiction linéaire (*LPC*) effectuée de manière synchrone à l'extraction des paramètres vidéos, c'est-à-dire toutes les 20 ms, sur des fenêtres de 32 ms, avec un recouvrement de 12 ms. On échantillonne d'abord sur ces enveloppes les valeurs de 32 canaux spectraux normalisés en énergie (on suppose qu'il n'y a pas de prédictibilité du niveau d'énergie d'un signal par la forme des lèvres). Puis, pour rendre plus aisée la modélisation statistique de la cohérence audiovisuelle, la dimension de l'espace audio est réduite de 32 à 5 par analyse en composantes principales, ce qui permet de conserver 85% de la variance des données. Au final, la dimension de l'espace audiovisuel est donc égale à 7 (2 dimensions vidéos + 5 dimensions audios).

Une première série de stimuli du corpus audiovisuel est utilisée à l'apprentissage du modèle statistique, et une deuxième série est réservée aux tests d'évaluation (séparation proprement dite). Chaque série représente $4 \times 6 \times 4 = 96$ stimuli possibles, soit environ 2300 vecteurs (environ 24 vecteurs par stimulus).

3.2. *Résultats*

Modélisation statistique

Les résultats de la modélisation statistique de la cohérence audiovisuelle sont présentés sur la Fig. 1. Huit gaussiennes ont été nécessaires pour modéliser le corpus d'apprentissage, c'est-à-dire représenter de manière correcte les probabilités des paires (spectres audio, paramètres labiaux). Les projections de ces gaussiennes dans le plan des deux dimensions visuelles et dans les trois premiers plans principaux de l'espace audio (Fig. 1) sont interprétables de la manière suivante.

Dans l'espace vidéo, on retrouve une organisation de base classique, avec les formes labiales fermées (bilabiales en contexte, gaussienne 1), arrondies ([y], [u] ainsi que dentales et vélaires dans ces contextes), gaussiennes 2, 3 et 8), étirées ([i], gaussienne 7) et ouvertes ([a], gaussienne 5). Les gaussiennes 4 et 6 modélisent les gestes d'ouverture-fermeture de la mâchoire et des lèvres entre ces cibles.

L'espace audio fait apparaître une propriété majeure des configurations audiovisuelles, déjà bien étudiée dans la littérature : la complémentarité des deux modalités [Rob98]. Ainsi, ce qui est proche visuellement est éloigné auditivement. Les gaussiennes 2, 3 et 8, quasiment confondues visuellement, se séparent bien auditivement, et l'on constate qu'elles correspondent respectivement aux catégories de type [y], [u] et [tdkg/yu] ; de même, les gaussiennes 5 [a] et 7 [i] sont très séparées par la première dimension spectrale. Au contraire, ce qui est proche auditivement est bien séparé visuellement : voir par exemple les gaussiennes 1, 4 et 8. Cette complémentarité est essentielle pour l'efficacité de notre méthode.

Débruitage

Nous avons donc mélangé la seconde répétition des séquences $V_1CV_2CV_1$ *non apprises* à des phrases prononcées par un autre locuteur, avec des coefficients de mélange m_{12} et m_{21} variant aléatoirement entre 0 et 2. Rappelons que l'algorithme consiste à calculer un signal $s_1=x_1-c_1x_2$ et à déterminer la valeur optimale de c_1 selon le principe suivant : lorsque c_1 varie, le spectre de s_1, soit S_1, décrit une courbe dans l'espace spectral à 5 dimensions, et on sélectionne le point de cette courbe maximisant $p(S_1, v_1)$. On voit dans la Fig. 2 le type de résultat obtenu : la valeur de c_1 optimale permet bien de retrouver le spectre original A_1 de façon très acceptable. On comprend pourquoi la complémentarité audiovisuelle est un ingrédient essentiel du succès : l'estimation de c_1 sera d'autant meilleure que les trajectoires du spectre audio $S_1(c_1)$ seront mieux alignées avec les lignes de plus grandes pentes de la probabilité $p(S_1, v_1)$. Lorsque la valeur de c_1 est bien estimée, le système réalise alors un débruitage quasiment parfait et rapide. On retrouve un signal s_1 très proche de a_1.

On peut par contre s'attendre à des difficultés lorsque la configuration visuelle est ambiguë, ce qui est, on le sait, souvent le cas. L'approche probabiliste prend ici tout son intérêt : si l'on est dans un cadre de mélange quasi-stationnaire, avec des valeurs à peu près stables des coefficients m_{ij} (approximation classique en séparation de sources), on peut estimer c_1 en optimisant une probabilité cumulée sur plusieurs trames consécutives :

$$c_1 = \arg\max\left(\Pi_t\ p\left(S_1(t), v_1(t)\right)\right)$$

Les configurations visuelles moins ambiguës peuvent alors guider peu à peu l'estimation de c_1. Nous avons pu ainsi, en utilisant un processus de recherche exhaustive de la valeur optimale de c_1, obtenir une estimation exacte (selon un critère d'erreur relative de 10%) dans 77% des cas (2300 estimations environ) sur une trame, et dans 88% des cas en cumulant la probabilité sur 6 trames consécutives. Ceci fournit un débruitage quasi-parfait dans le cas d'un mélange stationnaire sur chaque séquence.

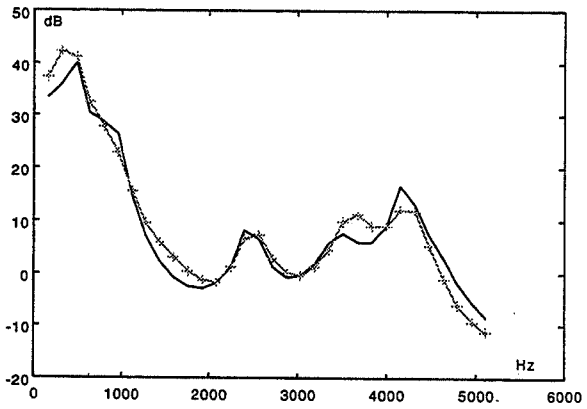


Figure 2 : Exemple de débruitage (son [u]): le spectre débruité S_1 (croix) réalisant le maximum de $p(S_1, v_1)$ est proche du spectre original A_1 (en foncé).

4. CONCLUSION

Le principe de séparation de sources audiovisuelles que nous avons introduit semble viable, avec un cadre probabiliste dont on sait qu'il est le bon pour résoudre ce genre de problème. Evidemment, nous n'avons pour le moment pas fait mieux que ce qui existe déjà : les algorithmes traditionnels de séparation de sources réussissent parfaitement à séparer des mélanges additifs stationnaires à 2 sources et 2 capteurs. Mais le caractère prometteur de ce travail est que l'on peut, sur ce principe, espérer travailler sur moins de capteurs que de sources, et utiliser l'information visuelle pour se focaliser sur une source et la débruiter plus efficacement et avec des temps de convergence plus rapide qu'avec les techniques audio pures... et c'est d'ailleurs manifestement ce qui se passe dans les situations de "*cocktail-party*" ! A partir de là, l'enjeu théorique est de marier techniques de séparation "informationnelles" et notre algorithme exploitant la cohérence multimodale, dans un même cadre qui sera fourni naturellement par la théorie de l'information : c'est là-dessus que porteront nos efforts à venir.

BIBLIOGRAPHIE

[All99] A. Allard, *Séparation de sources multimodales utilisant l'image du locuteur parlant*, INPG, DEA Signal-Image-Parole, 1999.

[Com94] P. Comon, Independent Component Analysis, a new concept ?, *Signal Processing*, Elsevier, 36(3):287-314, 1994,. Special issue on Higher-Order Statistics.

[Gir98] L. Girin, L. Varin, G. Feng & J.L. Schwartz, A signal processing system for having the sound "pop-out" in noise thanks to the image of the speaker's lips: new advances using multi-layer perceptrons, *Proc. 5th Int. Conf. Spoken Language Proc. (ICSLP'98)*, Sydney, 1998.

[Sto96] D.G. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines: theories, models and applications*, Springer-Verlag, Berlin, 1996.

[Rob98] J. Robert-Ribes, J.L. Schwartz, T. Lallouache, & P. Escudier, Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise, *J. Acoust. Soc. Am.*, 103:6, pp. 3677-3689, 1998.

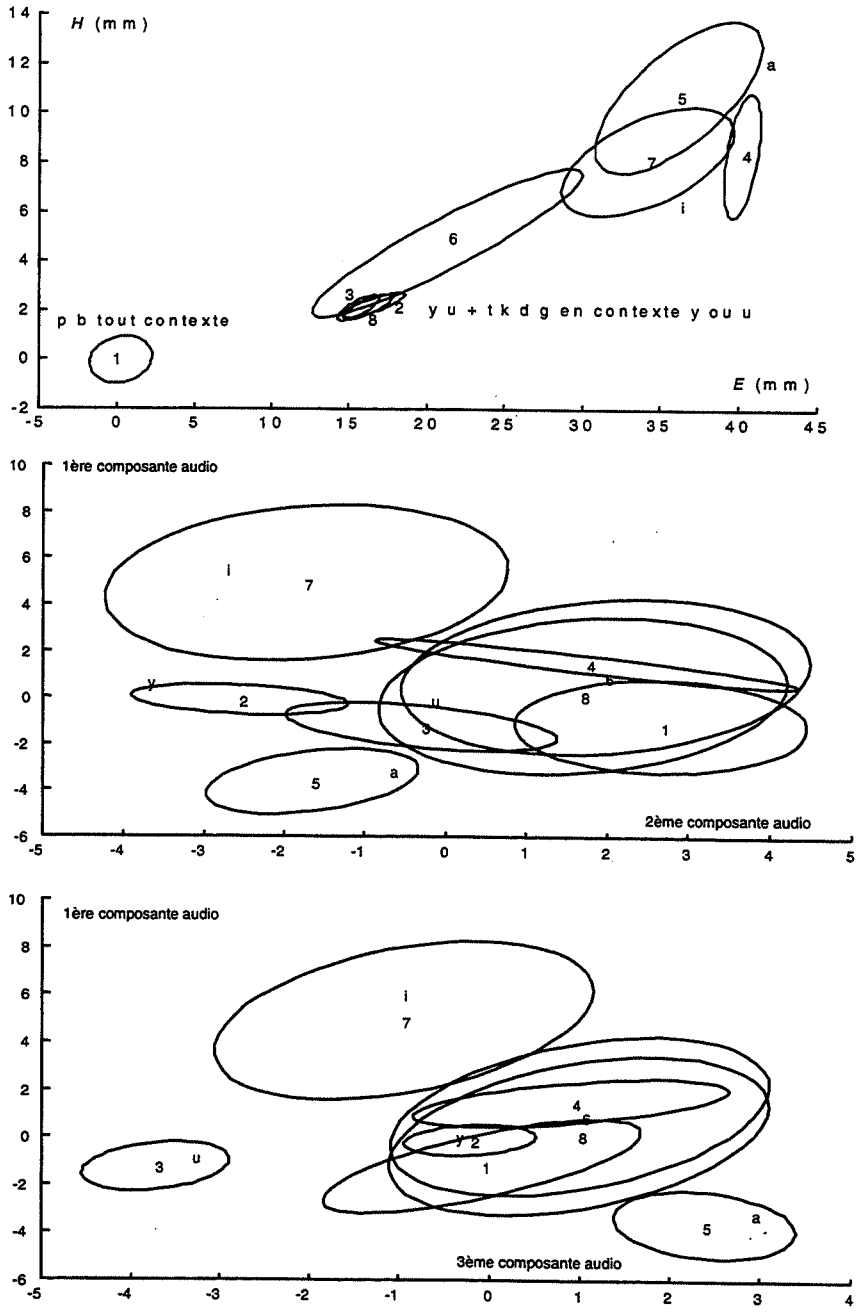


Figure 1 : Projection des ellipses de dispersion des gaussiennes dans le plan vidéo (*E, H*) et dans les trois premiers plans principaux audios. Les valeurs de *E* et *H* égales à zéro pour [p b] ont été modifiées à des valeurs aléatoires très faibles pour permettre la modélisation. La position typique des 4 voyelles [i], [a], [u] et [y] en position finale (peu réduite) est indiquée.

Reconnaissance de la parole dans le bruit après renforcement fondé sur l'harmonicité

Frédéric BERTHOMMIER et Hervé GLOTIN

Institut de la Communication Parlée/INPG
46, Av. Félix Viallet, 38031 Grenoble CEDEX
{bertho, glotin}@icp.inpg.fr

ABSTRACT

We propose and test a technique for speech enhancement based on the computation of a harmonicity index, which is related to the SNR. To carry out the performance evaluation, we quantify the accuracy of reconstruction of the target speech source. We vary factors including the size of the time-frequency regions in which the enhancement process is applied and the use of demodulation. We conclude that these factors have little effect on reconstruction accuracy, but demodulation improves the reconstruction and a process applied in 4 sub-bands with 128 ms time frame-duration is satisfactory. Then, using a HMM/ANN model, we evaluate the recognition scores, in comparison with those obtained with unprocessed noisy speech, J-RASTA-PLP pre-processing and training with a clean signal. A gain of 3-4dB is observed in loud noise with GWN, and 3dB with car noise, at WER=65%. We obtain the best gain after training with clean processed speech, but a significant gain is also obtained without such training.

1. INTRODUCTION

Les auditeurs humains sont aptes à identifier la parole dans un bruit fort, dans des conditions de bruitage variées (stationnaires ou non), ainsi que dans des conditions d'interférence avec un second locuteur. Des expérimentations psycho-acoustiques ont permis de caractériser l'effet de "flux auditif", qui est la perception de sources bien séparées et organisées comme un ensemble "d'objets auditifs". La modélisation de ce phénomène motive l'approche CASA (en Français: Analyse de Scène Auditive Computationnelle). Une hypothèse formulée dans ce champ de recherche est que ce phénomène résulte d'un traitement auditif des sons complexes fondé sur l'analyse de caractéristiques primitives telles que l'harmonicité et la localisation spatiale. Une seconde hypothèse est que cette analyse contribuerait à la robustesse de l'identification des sources, en particulier de la parole, soit directement, comme nous allons le montrer, soit à travers le processus d'organisation auditive, comme nous le suggérons également.

L'approche classique adoptée pour améliorer la robustesse de la reconnaissance de la parole par la machine est fondée sur l'optimisation du pré-traitement visant à extraire, puis à paramétrer les caractéristiques propres du signal de parole. Par conséquent, des attributs primitifs qui seraient à la base de la robustesse de la perception humaine (et de la construction des flux auditifs), ne sont pas pris en compte. Les performances de ces systèmes sont rapidement dégradées dans un bruit fort et nous pouvons espérer une amélioration en incorporant d'autres sources d'information.

Dans cet article, nous proposons un modèle d'extraction et d'usage de l'information de voisement apparenté à un filtrage de Wiener, et que nous nommons aussi "CASA front-end" en Anglais pour sa propriété de ségrégation du signal et du bruit. L'effet obtenu est une séparation incomplète du signal de parole et du bruit interférant à condition que celui-ci soit inharmonique. Le principe de renforcement est celui du filtrage de Wiener: dans le domaine fréquentiel, un avantage est donné aux composantes pour lesquelles le signal prédomine sur le bruit, et ainsi, le signal est renforcé globalement par rapport au bruit.

2. DESCRIPTION DU MODELE

Nous utilisons une quantification de l'harmonicité du signal liée au voisement, l'indice R, afin d'estimer localement dans le plan temps-fréquence le niveau de bruitage de la source désirée (voir [Gro00], ce Vol.). Cette valeur dépend du rapport signal sur bruit (SNR en Anglais), mais n'en est pas une estimation directe. Plusieurs méthodes ont été proposées pour estimer le SNR, directement dans le domaine fréquentiel [Kli87], ou bien, plus récemment, dans le domaine temporel à partir de l'autocorrélation [Boe93]. Nous avons également développé une méthode de marquage de la représentation temps-fréquence fondée sur l'usage du même indice ([Ber98],[Ber99]), qui est compatible avec les modèles de reconnaissance dits "multi-bandes" [Bou96]. Nous proposons ainsi plusieurs solutions complémentaires pour introduire l'information de voisement dans le processus de reconnaissance de la parole (voir aussi [Gai99]).

2.1 La représentation temps-fréquence

Nous appliquons sur le signal d'entrée un filtrage par banc de filtres fondé sur la FFT. La fréquence d'échantillonnage est de 8 kHz. Afin de diviser le plan temps-fréquence en régions de taille variable, nous faisons appel à la méthode de décomposition proposée par Tessier et coll. [Tes99]. L'intérêt principal de cette méthode est de contrôler la taille des régions temps-fréquence en faisant varier le nombre de filtres nc=(4,8,16) et la durée de chaque trame temporelle entre (512, 1024, 2048) échantillons.

Pour chaque filtre Fi, nous définissons une fenêtre de Hanning dans l'échelle Bark, centrée sur Fci, puis nous replaçons cette fenêtre dans l'échelle linéaire de la FFT. La conversion Hz/Bark est donnée par [Sha87]:

$$F_{Bark} = 13\ \mathrm{atan}(\frac{0.76\,F_{Hz}}{1000}) + 3.5\ \mathrm{atan}(\frac{F_{Hz}^2}{7500^2})$$

Les fréquences centrales Fci sont calculées à partir des bornes Fmin et Fmax du domaine couvert par le banc et filtres et à partir du nombre de filtres nc (figure 1). L'intervalle séparant deux filtres est égal à Fint=(Fmax-Fmin)/nc. Les deux filtres extrêmes ne couvrent que 1.5 intervalle tandis que les nc-2 filtres centraux s'étendent sur 2 intervalles. La fréquence centrale des filtres est Fci=Fmin+(i-0.5)*Fint.

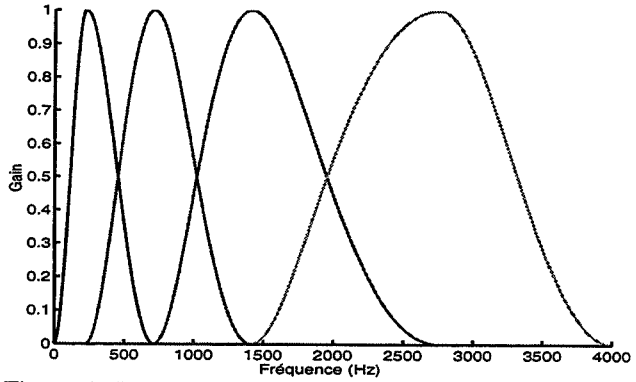


Figure 1: Banc à 4 filtres. Entre 0 et 4000 Hz, nous avons Fint=4.31 Bark, et Fci=[2.16, 6.47, 10.79, 15.10]Bark.

Pour chaque trame du signal d'entrée n, les spectres sous bandes obtenus sont:

$$|X_{n,i}(\omega)| = F_i(\omega).|X_n(\omega)|$$

2.2 Calcul et usage de l'indice R

Nous allons comparer plusieurs versions de notre modèle. D'une part, nous utiliserons deux versions de l'algorithme d'évaluation de l'indice R, l'une faisant appel à la démodulation du signal en sous-bandes (appelée proc1) et l'autre pas (proc2). Pour proc1, le signal sous-bandes re-synthétisé par FFT inverse (iFFT) est démodulé. Cette étape consiste en une rectification simple alternance suivie d'un filtrage passe-bande trapézoïdal [0,90,350,1000]Hz. Pour proc1 et proc2, l'index évalué est le rapport R= R1/R0. R1 est la valeur maximale de l'autocorrélogramme prise dans une fenêtre d'observation 1/[350,90]s et normalisée par l'amplitude R0 du pic en zéro. C'est la racine carrée de R qui est utilisée pour calculer le facteur de pondération W:

$$\begin{cases} \text{if } (R_{1,i} / R_{0,i}) > 0 \text{ then } R_i = (R_{1,i} / R_{0,i}) \\ \text{else } R_i = 0 \\ W_i(R_i) = R_i^{0.5} \end{cases}$$

L'index d'harmonicité R est une estimation du rapport (énergie périodique/énergie totale) elle-même en relation avec le rapport signal/bruit à condition que le signal possède des composantes périodiques et que le bruit soit apériodique. Le choix de l'exposant 0.5 est motivé par le fait que nous pondérons le spectre d'amplitude, mais aussi par l'existence d'un compromis entre le niveau de renforcement et le niveau de distorsion des signaux résultants. Le facteur de pondération W permet d'évaluer le spectrogramme de la source c, de même que celui de la source interférente n à partir de (1-W), ce qui correspond à la notion de "fonction de partage" et de ségrégation des sources [Tes99]. Le facteur W est appliqué sur le spectrogramme au niveau des régions temps-fréquence:

$$|\hat{X}_{c,i}(\omega)| = W_i(R_i).|X_{n,i}(\omega)|$$

Remarquons que l'effet obtenu n'est pas une ségrégation complète des composantes du signal à cause de l'application par sous-bande. Le spectre reconstruit de chaque fenêtre temporelle est égal à la somme des spectres sous-bandes:

$$|\hat{X}_c(\omega)| = \sum_{i=1}^{nc} |\hat{X}_{c,i}(\omega)|$$

Au total, la structure spectro-temporelle fine du signal est préservée afin de réaliser la re-synthèse par iFFT.

3. EVALUATION DU MODELE

Lorsque les signaux sont re-synthétisés, nous quantifions leur distorsion par rapport aux signaux clairs à l'aide d'une mesure de similarité entre les spectres. Le degré d'atténuation du bruit est aussi appréciable à l'écoute. Puis le gain du renforcement est évalué à partir des taux de reconnaissance automatique obtenus pour différents niveaux de bruit, par comparaison entre signaux non traités (proc0) et signaux traités (proc1 et proc2). Le modèle de reconnaissance utilisé comporte une étape de pré-traitement qui lui est propre (J-RASTA-PLP, [Her94]) et le gain que nous mesurons est cumulatif par rapport à celui conféré par cette méthode, qui est également efficace contre les bruits stationnaires que nous utilisons.

3.1 Evaluation directe à partir du signal reconstruit

Nous utilisons le signal clair c de module |Xc| afin de mesurer la précision de la reconstruction obtenue à partir du signal bruité n, en faisant appel à une référence. Nous définissons le RA, proposé par [Yan92], dans le domaine spectral. Nous complétons cette estimation à l'aide du SNRI, qui prend également en compte le signal bruité. Dans chaque trame temporelle, tous les spectres "pleine bande" sont renormalisés, de telle sorte que la somme des modules soit égale à 1:

$$RA = 10\log\frac{\int_{\Omega}|X_c(\omega)|^2}{\int_{\Omega}(|X_c(\omega)|-|\hat{X}_c(\omega)|)^2}$$

$$SNRI = 10\log\frac{\int_{\Omega}(|X_c(\omega)|-|X_n(\omega)|)^2}{\int_{\Omega}(|X_c(\omega)|-|\hat{X}_c(\omega)|)^2}$$

où Ù/2ð =[0,4000]Hz

Une statistique de RA et SNRI est établie pour toutes les trames, silences inclus, des mêmes 100 phrases de la partie "test" de NB95 (base multilocuteur de "digits" téléphonés, à 8kHz). La durée de la trame d'analyse est fixée à 1024 échantillons, proche de celle utilisée pour les tests de reconnaissance (1000 éch.), avec recouvrement de moitié. L'effet des deux facteurs **(1)** nombre de sous-bandes (nc), et **(2)** longueur de la fenêtre de traitement, est analysé en additionnant un bruit blanc gaussien (GWN) à 0dB (table 1). Le facteur nc a un petit effet négatif pour proc2, et le facteur durée présente un petit effet positif pour les deux. Nous voyons que proc1 est légèrement meilleur que proc2, mais il n'y a pas de différence observée à 0 dB pour la condition principale de l'étude, qui correspond à celle des tests de reconnaissance (nc=4, 1024 éch., soit 128 ms).

nc/éch.	512	1024	2048
4	6.1/6.0	**6.4/6.3**	6.5/6.4
8	6.1/5.8	6.5/6.2	6.7/6.4
16	6.0/5.3	6.3/5.6	6.5/5.8

Table 1: Moyenne de RA en dB pour proc1/proc2, sur toutes les trames de 100 phrases de la base de test. Ligne: variation de la durée. Colonne: variation du nombre de sous-bandes (nc). Le SNRI moyen (non figuré) est bien corrélé avec le RA.

Ensuite, pour cette condition, nous faisons varier le niveau du bruit blanc entre –18 et 21 dB (figure 2). Le SNR est ici exprimé en dB RMS silence inclus relativement au signal clair c. La figure 3 montre que proc1 est meilleur que proc2 lorsque le SNR est élevé, au dessus de 0dB, mais pas au dessous. De plus, nous observons que la courbe de SNRI est non monotone et présente un maximum à environ 6dB.

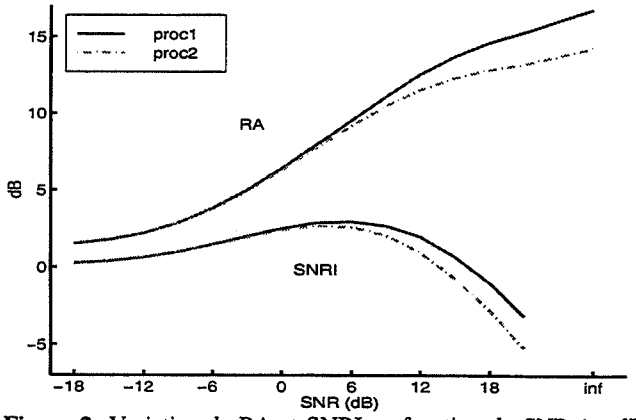


Figure 2: Variation de RA et SNRI en fonction du SNR (en dB RMS, avec du bruit GWN). INF est obtenu avec le signal clair.

3.2 Tests de reconnaissance

La source désirée est reconstruite afin d'alimenter un système de reconnaissance pleine bande de type ANN/HMM. Nous testons la possibilité d'adapter le processus de reconnaissance à ce mode de renforcement au cours de la phase d'apprentissage réalisée à partir des signaux clairs. Ceci est motivé par le fait que des distorsions sont introduites par ce traitement car il favorise les signaux voisés au détriment des signaux non voisés.

Nous entraînons le modèle dans 3 conditions différentes i=0,1,2 à partir des signaux d'origine de NB95 (proc0), ou bien à partir des signaux renforcés proc1 et proc2. Les paramètres sont fixés à nc=4 et durée=1024. Trois perceptrons multicouches (mlp0, mlp1, mlp2) résultent de cette phase d'adaptation. La méthode de pré-traitement des signaux d'entrée proci est toujours un J-RASTA-PLP afin de rechercher un effet coopératif entre l'extraction des caractéristiques propres au signal de parole et l'effet de renforcement lié à l'harmonicité.

Pour chaque point de la phase de test, nous ferons toujours appel aux mêmes 100 phrases de la base de test de NB95. Nous noterons proci>mlpj le test du reconnaisseur mlpj (entraîné avec des signaux procj) avec un signal d'entrée proci. Les performances sont établies en WER ("Word Error Rate", en Anglais). Nous appliquons tout d'abord ce test sur les signaux obtenus en faisant varier les paramètres de durée et nc (ceux du §3.1) pour observer table 2 que le RA n'est pas bien corrélé au WER dans plusieurs conditions. Nous voyons à présent que proc2>mlp2 est meilleur que proc1>mlp1.

nc/éch.	512	1024	2048
4	38/35	**38/34**	41/40
8	38/34	39/35	40/38
16	48/41	49/42	50/39

Table 2: %WER moyen à 0dB pour proc1>mlp1/proc2>mlp2, sur 100 phrases de la base de test. Ligne: variation de la durée de la trame (en échantillons) pour l'application de proci. Colonne: variation du nombre de sous-bandes (nc). Le taux d'erreur (WER) de proc0>mlp0 est de 47% à 0dB.

Puis nous fixons les paramètres à nc=4 et durée=1024 éch. pour établir une relation entre WER et SNR avec du bruit blanc gaussien (GWN). Le RA est corrélé négativement avec le SNR, et nous observons aussi que proc2>mlp2 est légèrement meilleur que proc1>mlp1 (figure 3). Nous évaluons le gain de performance des méthodes de renforcement proc1 et 2 par comparaison avec proc0>mlp0 faisant uniquement appel au J-RASTA-PLP comme méthode de pré-traitement. Ce gain correspond à une mesure du décalage exprimé en dB au point WER=65% placé au centre de l'intervalle [30-100]%, au niveau duquel nous observons une différence significative. Le gain est significatif au dessous de 3-6dB, sans perte importante pour le signal clair.

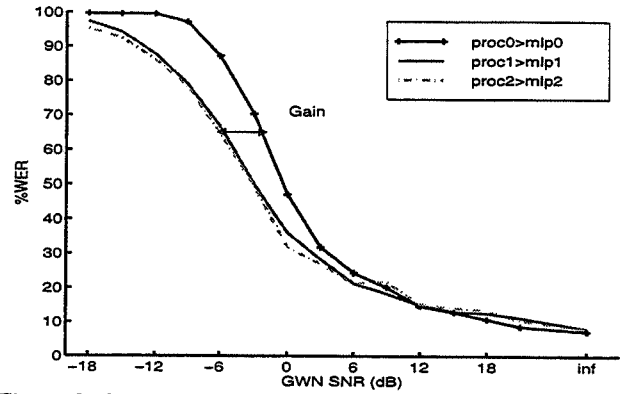


Figure 3: Courbe de réponse du modèle en %WER avec du bruit blanc GWN en comparaison avec la référence proc0>mlp0, et calcul du gain |ΔWER65|. Le %WER de proc0>mlp0 en clair (INF) est de 7.3%.

Un second test est réalisé à l'aide d'un bruit stationnaire de voiture (véhicule roulant à 80 km/h fenêtres fermées), dans des conditions strictement équivalentes (figure 4). Les performances globales sont similaires à celles observées pour le bruit blanc GWN et les combinaisons proc1>mlp1 et proc2>mlp2 sont équivalentes. Mais nous constatons un gain plus faible et une petite dégradation des performances au dessus de 3dB pour proc1>mlp1 et proc2>mlp2. Cela est probablement dû à l'atténuation des phonèmes non voisés.

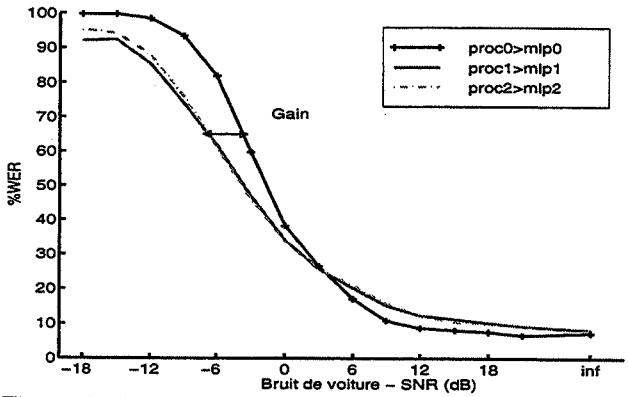


Figure 4: %WER obtenu en faisant varier le niveau d'un bruit stationnaire de voiture.

Enfin, nous établissons le gain pour les combinaisons proci>mlpj avec les deux types de bruit (table 3).

	GWN	Bruit de voiture
mlp0	3.0/3.0	1.9/1.9
mlp1	3.4/4.0	3.1/3.1
mlp2	3.2/3.7	3.0/3.0

Table 3: |ΔWER65| de proc1/proc2 en dB, pour le bruit blanc GWN et le bruit de voiture, obtenu avec les différent mlpi. Les points de référence appartiennent à la fonction proc0>mlp0 et ont pour valeurs -2.3dB (GWN) et -3.8dB (bruit de voiture) en WER=65%.

4. CONCLUSION

Nous montrons une amélioration significative des taux de reconnaissance dans des bruits stationnaires forts à partir d'une méthode de renforcement utilisée conjointement avec le pré-traitement de type J-RASTA-PLP. Les segments vocaliques sont les plus résistants aux bruits forts par leur intensité globale et parce que les trajectoires formantiques sont à la fois saillantes, redondantes et continues temporellement. L'effet du renforcement est cumulatif dans ces conditions. Par rapport à une méthode spectrale [Kli87], l'usage d'un algorithme d'évaluation temporelle de l'indice d'harmonicité est attrayant puisqu'il est **(1)** simple et rapide, **(2)** il fait appel à peu de connaissances a priori, et **(3)** il est plausible en tant que modèle auditif [Gro00]. Nous montrons aussi que la démodulation (proc1), ainsi que l'adaptation (mlp1 et 2) sont optionnelles. L'évaluation de l'index d'harmonicité étant locale spectralement et temporellement, nous prévoyons également un gain important avec des bruits non stationnaires.

Remerciements: Ce travail est réalisé dans le contexte des projets Européens TMR SPHEAR et LTR RESPITE. Nous remercions Fritz Class (Daimler-Benz) pour nous avoir donné un ensemble de bruits de voiture. Les tests de reconnaissance de la parole ont été réalisés à l'IDIAP à partir du logiciel STRUT développé au FPMs Mons, et de la base de données OGI Numbers95 (NB95 dans le texte).

BIBLIOGRAPHIE

[Ber98] Berthommier, F., Glotin, H., Tessier, E., Bourlard, H. (1998) Interfacing of CASA and partial recognition based on a multistream technique, ICSLP'98, Sydney, pp. 1415-1418.

[Ber99] Berthommier, F., Glotin, H. (1999) A new SNR-feature mapping for multistream speech recognition, ICPHS'99, San Francisco, vol. 1, pp. 711-714.

[Boe93] Boersma, P. (1993) Accurate short-term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound, IFA Proc, Amsterdam, 17:97-110.

[Bou96] Bourlard, H., Dupont, S., Hermansky, H., Morgan, N. (1996) Towards subband-based speech recognition, EUSIPCO, pp. 1579-1582.

[Gai99] Gaillard, F., Berthommier, F., Feng, G., Schwartz, J-L. (1999) A reliability criterion for time-frequency labelling based on periodicity in an auditory scene, Eurospeech'99, Budapest.

[Gro00] Grosgeorges, A., Berthommier, F., Apoux, F., Lorenzi, C. (2000) Détection de la modulation d'amplitude liée au voisement: comparaison entre expérimentation et modélisation, JEP 2000 (ce Vol.).

[Her94] Hermansky, H., Morgan, M. (1994) RASTA processing of speech, IEEE Trans. on Speech and Audio Processing, 2:4:578-589.

[Kli87] Klingholz, F. (1987) The measurement of the signal to noise ratio (SNR) in continuous speech, Speech Com., 6:15-26.

[Sha87] O'Shaughnessy, D. (1987) Speech communication: Human and Machine, Addison-Wesley, New-York.

[Tes99] Tessier, E., Berthommier, F., Glotin, H., Choi, S. (1999) A CASA front-end using the localisation cue for segregation and then cocktail-party speech recognition, ICSP'99, Seoul, pp. 97-102.

[Yan92] Yang, X., Wang, K., Shamma, S., A. (1992) Auditory representations of speech signals, IEEE Trans. on Inf. Theory, 38:2:824-839.

Indexation de la bande sonore : les composantes Parole/Musique

Ludovic Fontaine, Christine Sénac, Nathalie Vallès-Parlangeau, Régine André-Obrecht

Institut de Recherche en Informatique de Toulouse
UPS – 118, Route de Narbonne – Toulouse, France
Tél.: ++33 (0)561558835 - Fax: ++33 (0)561556258
Mél: {fontaine, senac, parlange, obrecht}@irit.fr - http://www.irit.fr/

Abstract

This work addresses the soundtrack indexing of multimedia documents. We aim at isolating music and speech parts. A first study of an indexing system based on a differentiated modelling is presented: each class is modelised and has its own model and representation space. Experimentations are conducted on an episod of the television serial "Chapeau melon et bottes de cuir" ("The avengers"). The first results are very satisfactory in respect of the nature of the corpus and the limited quantity of training data.

1. Introduction

Le volume de données numériques actuellement disponible est en très nette augmentation (librairies audio sur Internet, bases de données, bouquets numériques,...). Leur exploitation est en grande partie conditionnée par leur facilité d'accès. Aussi, développe-t-on des modalités d'accès dites 'intelligentes' dans le cadre d'applications diverses dans des domaines tels que les télécommunications, l'éducation, l'expertise, l'archivage, la télévision, le cinéma... que ce soit pour des applications grand public ou professionnelles. Les contenus multimédia de ces documents sont des images fixes, de la vidéo, des bandes sonores ou encore du texte. L'accès à l'information contenue dans les différentes composantes d'un document multimédia est étroitement lié à leur indexation.

L'indexation de la bande sonore d'un document multimédia consiste en une indexation par le contenu [Gau99]. Citons par exemple la recherche de 'bruits' ou de sons sémantiquement significatifs tels que les applaudissements ou les effets spéciaux (explosions,...). D'autres informations pertinentes peuvent être la détection de locuteurs signifiant des tours de parole dans un dialogue, ou encore leur identification s'ils sont connus *a priori*. La recherche de mots clés (mots isolés, groupes de mots,...) est une information importante sur le contenu du message verbal. Toutes ces études supposent la distinction Parole/Musique/Bruit en amont. Elle permet non seulement de fournir une indexation par son contenu musique, mais aussi de circonscrire une zone de recherche plus précise pour toutes les informations précédemment citées.

Parmi les méthodes de discrimination Parole/Musique classiquement trouvées dans la littérature, nombre de chercheurs se sont intéressés aux différences acoustiques qui peuvent exister entre ces deux types de sons. [Sau96], [Sch97], [Par99], [Car99] ont basé leur discrimination sur un ensemble plus ou moins important d'indices tels que le taux de passages à zéro, la variation de flux spectral, des mesures de rythmicité, ... Les méthodes de classification restent plus classiques (Modèles de Mélanges de lois Gaussiennes, k plus proches voisins,...). D'autres ont gardé une paramétrisation de type cepstrale associée à des Modèles de Mélanges de lois Gaussiennes et montrent des résultats satisfaisants [Sec99].

Nous proposons ici une première étude d'un système d'indexation Parole/Musique basée sur une modélisation différenciée pour chacune des classes parole et musique. L'approche est mise en œuvre à partir de Modèles de Mélanges de lois Gaussiennes La première partie met en perspective l'intérêt de la modélisation différenciée et son application dans le cadre de cette tâche d'indexation. La mise en œuvre du système et les expérimentations sont ensuite détaillées, avant de donner un bilan positif de cette approche.

2. La modélisation différenciée

2.1 La décomposition Parole/Musique

Dans une approche classique de discrimination Parole/Musique, il est question d'un choix binaire : parole ou musique. Dès lors qu'il s'agit d'indexation, le but est de trouver les composantes parole et musique du document sonore de façon indépendante.

On recherche dans le document les parties contenant de la parole (resp. de la musique). Ces parties seront annotées 'Parole' (resp. Musique) contrairement aux parties ne contenant pas de parole (resp. musique) annotées 'NonParole' (resp. NonMusique). Le document sera donc annoté de deux manières indépendantes : l'une pour la parole et l'autre pour la musique. Il n'est donc plus question de chercher à discriminer la parole de la musique, mais à les caractériser au mieux de façon indépendante afin de faire une séparation de type Classe/NonClasse (c'est-à-dire Parole/NonParole et Musique/NonMusique).

2.2 Le modèle théorique

Classiquement, en discrimination, les classes que l'on cherche à séparer partagent à la fois le même espace de représentation et la même modélisation. Dans la situation présente, les différences de production qui peuvent exister entre parole et musique se retrouvent tout naturellement dans la nature des signaux eux-mêmes : la parole se caractérise par une structure formantique, tandis que la musique se caractérise par une structure harmonique. Le but n'est donc plus de trouver des paramètres qui permettent de séparer au mieux ces classes, mais plutôt de trouver des ensembles de représentations qui caractérisent au mieux chacune des classes. De même, il peut être nécessaire de mettre en œuvre des modélisations distinctes pour chacune des classes.

La modélisation différenciée est donc tout à fait adaptée à notre problème ; elle est basée sur le principe que les différentes classes peuvent être modélisées séparément afin de prendre en compte leurs spécificités. Ainsi, chacune des classes est définie par son espace de représentation et son modèle.

1 classe = { Espace de représentation, Modèle}

2.3 Le système d'indexation

Le système est composé de deux modules : le prétraitement acoustique et la modélisation. La décomposition Parole/Musique se faisant de façon disjointe et sur le modèle Classe/Non Classe, ces deux modules sont totalement distincts pour chacune des deux classes. Ils sont suivis d'un éventuel module de fusion qui doit être développé en fonction de l'application réelle visée. Soit cette fusion peut être faite lors de l'indexation ; un choix doit alors être fait, impliquant parfois des confusions ou des pertes d'informations. Soit la fusion est propre au système de recherche d'informations exploitant l'indexation et auquel cas la sortie de l'indexation est constituée de deux flux totalement distincts. Dans l'application réelle visée ici, cette fusion est faite lors de la recherche d'informations, et ce module n'est donc pas mis en place.

3. Mise en oeuvre

3.1 Prétraitements acoustiques

Dans le cadre de la modélisation différenciée, deux prétraitements acoustiques ont été développés. Le premier a consisté en une analyse cepstrale selon une échelle Mel, effectuée sur des trames de 10ms. Au total 20 paramètres sont utilisés : 9 MFCC, l'énergie et les dérivées associées. Le second prétraitement a consisté à extraire 29 paramètres (28 sorties de filtres et l'énergie). La répartition des filtres dans le domaine spectral est linéaire par morceaux.

3. 2 Modélisation

La modélisation est basée sur un mélange de lois Gaussiennes. L'initialisation des modèles est obtenue par Quantification Vectorielle basée sur l'algorithme de Lloyd. L'étape d'optimisation des paramètres est réalisée par l'algorithme classique Expectation-Maximization (EM). L'indexation de chacune des classes est basée sur deux MMG distincts permettant la discrimination Classe/NonClasse. Après expérimentation, le nombre de lois gaussiennes a été fixé à 32 pour les modèles Parole et NonParole et à 10 pour les modèles Musique et NonMusique. Les matrices de covariances sont diagonales. A l'issue de l' indexation, certaines insertions de segments de taille négligeable sont à noter pour chacune des classes. Une procédure de lissage permet aux segments adjacents de les absorber.

4. Expérimentations et Résultats

4.1 Corpus

Le corpus est un épisode télévisé de la série 'Chapeau Melon et Bottes de Cuir', échantillonné à 16kHz. La durée totale du corpus est de 50 minutes. C'est un premier objet d'expérimentation qui a l'avantage de présenter de longues périodes de parole comme de musique et surtout des zones dites 'mixtes' contenant de la parole et de la musique et/ou du bruit,... Le corpus comporte de la parole dans diverses conditions plus ou moins bruitées (parole téléphonique, enregistrements en extérieur, poursuites en voiture, foule...) pour environ cinq locuteurs principaux (1 femme et 4 hommes). La musique est de la musique de variétés : cordes (très peu utilisées), vents (surtout les cuivres) et le trio basse, guitare électrique (surtout wha-wha) et batterie-percussions. Quelques autres instruments (harpe, piano,...) sont également utilisés sporadiquement.

Pour les besoins de l'expérimentation, trois annotations manuelles indépendantes ont été effectuées pour la parole, la musique et le bruit. L'étape ultime a consisté à regrouper ces trois annotations afin de générer une indexation qui comporte à la fois des segments dits ' purs ' (Parole, Musique ou Bruit) et des segments dits ' mixtes ' (table 1).

Table 1: Répartition des différents types de segments dans le corpus en terme de pourcentage de durée.

	%
Parole	24.4
Musique	23.1
Bruit	22.8
Parole/Musique	5.1
Parole/Bruit	5.6
Musique/Bruit	10.4
Parole/Musique/Bruit	8.6

4.2 Expériences et Résultats

Le corpus, d'une durée totale de 50mn a été découpé en deux parties. Sur la première, représentant 35mn de film, les segments de parole 'pure' (qui représentent 8'33 de parole) ont été extraits et ont servi à l'apprentissage du modèle Parole, les autres segments ont servi à l'apprentissage du modèle NonParole. De même, les segments de musique 'pure'(qui représentent 8'5 de musique) ont servi à l'apprentissage du modèle Musique alors que les autres segments ont servi à l'apprentissage du modèle NonMusique. Les tests ont été effectués sur la seconde partie du corpus (soit 15mn de film) qui contient évidemment les différents types de segments (segments 'purs' et segments 'mixtes'). Il est à noter que certains passages du corpus de test ne sont pas représentés dans le corpus d'apprentissage (nouveau locuteur, nouvelle musique, bruits d'ascenseur superposés à de la voix, parole téléphonique...).

L'évaluation de l'indexation automatique a été effectuée en comparaison avec l'indexation manuelle. Après avoir aligné les deux segmentations, nous avons mesuré différents délais entre les frontières automatiques et les frontières manuelles correspondantes si elles existent. Nous avons aussi noté le nombre d'insertions et d'omissions.

Parole

Seuls les résultats issus de l'analyse cepstrale sont détaillés, en effet les expériences issues de l'analyse spectrale linéaire n'ont pas donné de résultats satisfaisants pour la parole. Les modèles mis en œuvre sont des MMG à 32 gaussiennes avec un vecteur de 20 paramètres. Le corpus de test (15') comprend 291 segments manuels. L'évaluation est présentée ci-dessous (table 2).

Table 2: Evaluation de l'indexation de la Parole. Nombre des délais inférieurs à 20, 40 et 100 cs. Nombre d'omissions et d'insertions.

	<20cs	<40cs	<100cs	O	I
P	250	9	6	14	54
NP				12	7

Les résultats obtenus sont excellents: pour chaque frontière automatique, il existe toujours une frontière manuelle lui correspondant (même type de transition). Aucune substitution de segment n'a donc été relevée. Les délais sont tout à fait satisfaisants : 86% sont inférieurs à 20 centisecondes. La précision de la reconnaissance de la Parole (accuracy) est de 95%. Les résultats ne présentant pas de substitution, l'accuracy a été calculée comme suit: (durée du corpus de test – durée des Insertions – durée des Omissions) / durée du corpus de test

Les omissions de Parole sont principalement des zones de parole 'mixtes' ou dégradées (voix chuchotées ou criées, bruit important, haut-parleur,...). Les insertions de Parole sont principalement des gazouillis d'oiseaux, bruits de cloche, musique faible. Ces différents types de segments ne sont pas ou très faiblement représentés dans l'ensemble d'apprentissage.

Musique

Nous présentons ici les expériences à l'issue des deux prétraitements acoustiques. Le modèle mis en œuvre est un MMG à 10 gaussiennes avec un vecteur de 20 paramètres pour l'analyse cepstrale et un vecteur de 29 paramètres pour l'analyse spectrale linéaire. Le corpus de test (15') comprend 80 segments manuels. L'évaluation est présentée ci-dessous (table 3 et table 4).

Table 3: Evaluation de l'indexation de la Musique après analyse cepstrale. Nombre des délais inférieurs à 20, 40, 100 et 220 cs. Nombre d'omissions et d'insertions.

	<20cs	<40cs	<100cs	<220cs	O	I
M	50	7	8	8	7	17
NM					0	31

Table 4: Evaluation de l'indexation de la Musique après analyse spectrale linéaire.

	<20cs	<40cs	<100cs	<220cs	O	I
M	73	3	4	0	0	20
NM					0	16

La table 4 montre l'amélioration nette des résultats apportée par l'analyse spectrale linéaire. Le taux des délais inférieurs à 20cs passe de 62.5% à 91%. De même, l'accuracy pour la Musique passe de 81.5% à 93%.

Les omissions de Musique, qui correspondent principalement à des zones de musique très faible entre des segments de parole, disparaissent dans la table 4. Les insertions de Musique correspondent principalement à des bruits de voiture, de crissements de pneus, d'explosions et d'ascenseur. Ils ne sont pas représentés dans l'ensemble d'apprentissage. Les insertions de NonMusique correspondent à des segments comportant de la musique très faible ou des maracas.

4.3 Cohérence des résultats

Nous présentons ci-dessous (figure 1) un exemple d'indexation Parole et Musique sur un extrait de signal d'environ 12 secondes pour lequel nous avons réalisé l'indexation Parole/NonParole (b) et l'indexation Musique/NonMusique (c). Nous avons ensuite regroupé ces résultats (d) qui ont pu être comparés à l'étiquetage manuel (a).

L'analyse des résultats de la ligne (d), montre que les frontières obtenues sont correctes. Le premier petit segment Parole/Musique (2ème segment (d)) peut paraître insolite : sur le segment de Parole précédent, la parole est

mixée avec une note de musique tenue et de niveau sonore très faible. Pendant un court instant de silence sur la parole, cette note devient plus audible. Le segment Musique suivant ($4^{ème}$ segment de la ligne (c)) est le démarrage d'un morceau de musique suivi presque immédiatement d'un fort bruit de foule. Sur la ligne (d), les segments notés – (c'est-à-dire NonParole et NonMusique) correspondent à des bruits de foule.

5. Conclusion

Nous avons présenté les premières expérimentations d'un système d'indexation Parole/Musique basé sur une modélisation différenciée. L'idée de la méthode est d'associer à chacune des classes son propre espace de représentation et sa propre modélisation. Cette approche est mise en œuvre à partir de MMG sur la base d'une analyse cepstrale pour la Parole et d'une analyse spectrale linéaire pour la Musique.

Les résultats d'indexation obtenus, que ce soit pour la parole ou la musique sont excellents compte tenu de la nature du corpus et le volume restreint des données en apprentissage. L'indexation est fiable: pour chaque frontière automatique, il existe toujours une frontière manuelle lui correspondant associée au même type de transition. Aucune substitution de segment n'a donc été relevée. Les erreurs d'insertion et d'omission ont pu être expliquées à l'écoute et à l'observation du signal. La parole et la musique ayant des structures différentes (formantique pour la parole et harmonique pour la musique), la modélisation différentiée (analyse cepstrale pour la parole et analyse spectrale linéaire pour la musique) est tout à fait adaptée. Elle permet de choisir le meilleur espace de représentation pour chacune des deux classes. L'analyse cepstrale sur la parole et la musique donne respectivement une précision de reconnaissance de 95% et de 81.5%. Une amélioration très nette est obtenue en utilisant une analyse spectrale linéaire pour la musique: la précision passe alors à 93%.

Bibliographie

[Car99] Carey M.J., Parris E.S., Lloyd-Thomas H., "A comparison of features for speech, music discrimination", ICASSP'99.

[Gau99] Gauvain J.L., Lamel L., Adda G., "Audio partitioning and transcription for broadcast data indexation", CBMI'99, pp. 67-73.

[Par99] Parris E.S., Carey M.J., Lloyd-Thomas H., "Feature fusion for music detection", Eurospeech'99, pp. 2191-2194.

[Sau96] Saunders J., "Real-time discrimination of broadcast Speech/Music", ICASSP'96, pp. 993-996.

[Sch97] Scheirer E., Slaney M., (1997), " Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator ", ICASSP'97, Munich, Vol. II, pp. 1331-1334.

[Sec99] Séck M., Bimbot F., Zudaj D., Delyon B., "Two-class segmentation for speech/music detection in audio tracks", Eurospeech'99, pp. 2801-2804.

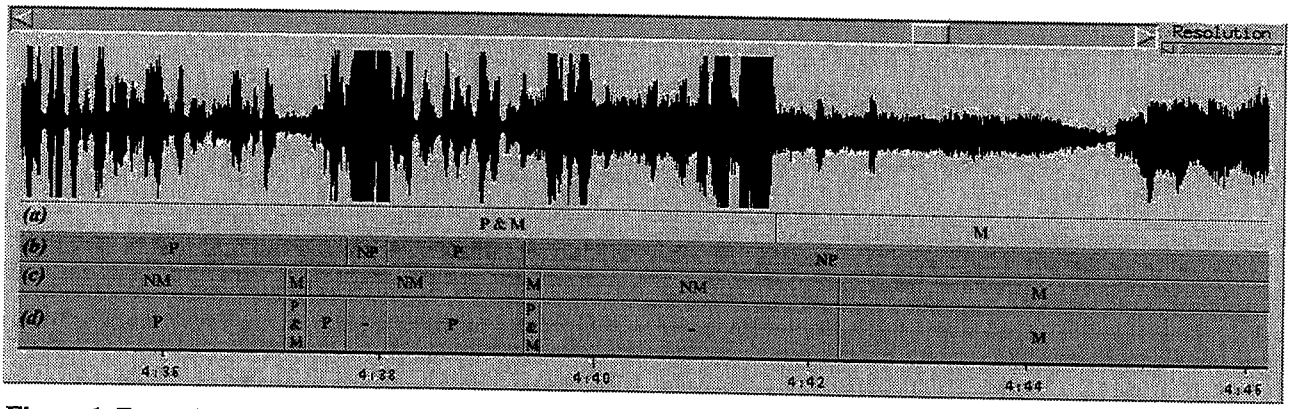


Figure 1: Exemple d'indexation Parole et Musique. (a) Etiquetage manuel (b) Indexation Parole/NonParole (P,NP) (c) Indexation Musique/NonMusique (M,NM) (d) Regroupement des indexations Parole/NonParole et Musique/NonMusique (P,M,P&M,-)

Modèle de Markov évolutif pour les tâches de suivi de locuteurs

Sylvain Meignier, Jean-François Bonastre, Corinne Fredouille, Teva Merlin*

LIA/CERI Université d'Avignon, Agroparc,
BP 1228, 84911 Avignon Cedex 9, France.
{sylvain.meignier, jean-francois.bonastre, corinne.fredouille, teva.merlin}@lia.univ-avignon.fr

Abstract

Seeking within a speech sequence the speaker utterances is one of the main tasks of indexing.

In this paper, the proposed speaker tracking system is defined in the case where all speaker identities are known beforehand. The conversation is modeled as an evolutive Markov Model, in which speaker models computed are added one by one. A temporary indexing process is proposed after each speaker adding and then challenged at the next step. This process is iterated until all the speakers are detected.

The system has been assessed using multi-speaker messages generated by concatenation of Switchboard mono-speaker segments. The obtained results show the potentiality of the proposed solution.

1. Introduction

La recherche au sein d'une conversation des paroles prononcées par différents locuteurs constitue une tâche essentielle pour l'indexation par le contenu de documents multimédia.

Deux approches usuelles, en indexation ou en suivi de locuteurs (*speaker tracking*), sont communément envisagées. La première méthode, décrite notamment dans [1] et [2], repose dans une première phase sur une détection des ruptures provoquées par les changements de locuteurs. Une seconde phase, dite de *clustering* détermine le nombre de locuteurs et groupe les segments par locuteur. Aucune information sur les locuteurs potentiels n'est utilisée pendant ces phases. Cette caractéristique rend la méthode bien adaptée aux tâches d'indexation en aveugle. La seconde méthode, proposée en particulier dans [3], est basée sur un système de reconnaissance du locuteur. La détection des segments et l'attribution de ceux-ci aux différents locuteurs sont réalisées simultanément. Dans cette approche, les locuteurs potentiels (et leurs modèles respectifs) doivent être connus par avance. Cette dernière contrainte destine particulièrement cette technique aux tâches de suivi de locuteurs.

Dans cet article, nous proposons un système apparenté à la deuxième méthode pour les tâches de suivi de locuteurs. Toutes les identités des intervenants sont connues *a priori*. La conversation est modélisée par un modèle de Markov évolutif (proche de celui proposé dans [3]). Au cours du processus d'indexation, le modèle évolue à chaque nouvelle détection d'un locuteur.

Le système proposé a été testé sur des messages générés à partir de la concaténation de segments de parole mono-locuteur issus de la base Switchboard (NIST 1998[1]). La base est constituée de segments de parole réelle, bruitée, provenant de conversations téléphoniques.

2. Modèle de conversation

2.1. *Structure du modèle de conversation*

La dynamique de la conversation est représentée par un modèle de Markov ergodique, où les états caractérisent les locuteurs du message et les transitions entre les différents états modélisent les changements de locuteurs.

Deux états particuliers sont introduits. L'état Parole détecte les blocs contenant des données devant être attribuées à un modèle de locuteur alors que l'état Non Parole indique les blocs contenant des silences, des perturbations importantes du canal de transmission, ou des bruits environnant les locuteurs enregistrés.

Le modèle est défini par :

- Un ensemble d'états $\{e_i\}_{1 \leq i \leq n}$ représentant les modèles de locuteurs.
- A chaque état est associé un ensemble de probabilités d'émission $\{a_{i,k}\}_{1 \leq i \leq n,\ 1 \leq k \leq m}$, calculées par un système de vérification du locuteur (en utilisant le modèle de locuteur correspondant à cet état, *cf* §*3*).
- Un ensemble $\{t_{i,j}\}_{1 \leq i \leq n,\ 1 \leq j \leq n}$ d'arcs entre les états.
- Un ensemble $\{b_{i,j}\}_{1 \leq i \leq n,\ 1 \leq j \leq n}$ de probabilités de transition associées aux arcs.

NB : Le modèle ergodique n'intègre pas de connaissance *a priori* sur le nombre et la durée des différentes interventions, ni sur la structure de la conversation.

* projet RAVOL : support financier du Conseil général de la région Provence Alpes Côte d'Azur et de DigiFrance.

[1] http://www.nist.gov/speech/spkrec98.html

2.2. *Calcul des probabilités de transitions*

Un ensemble de règles est utilisé pour définir la valeur des probabilités de transition. Cet ensemble est exprimé sous la forme d'une matrice de poids de passage d'un état à un autre.

Les poids de passage entre les états des locuteurs vérifient trois conditions :

- Les poids $p_{i,i}$ de rester dans le même état e_i sont égaux pour chaque état.
- Les poids $p_{i,j}$ entre deux différents états e_i, e_j sont égaux.
- Quelque soient e_i et e_j deux états différents alors $p_{i,j} < p_{i,i}$.

Les poids associés aux modèles Parole et Non Parole sont fixés par l'opérateur en fonction du coût d'une erreur d'indexation (bloc attribué à un mauvais locuteur) par rapport au coût d'une non décision (bloc non attribué à un locuteur).

Les poids sont reportés dans la matrice de transition, qui sera normée de sorte que la somme des valeurs des arcs sortant d'un état soit égale à 1.

2.3. *Construction du modèle par un processus itératif*

La construction du modèle de conversation est réalisée par un processus itératif, où les locuteurs sont détectés et ajoutés un à un au modèle de Markov.

A l'initialisation du processus (figure 1), le modèle de Markov est composé de deux états, qui représentent respectivement le modèle de Parole et le modèle de Non Parole.

En fin d'initialisation, l'algorithme de Viterbi, appliqué au modèle de Markov, donne une première segmentation, qui sera remise en cause à l'étape suivante.

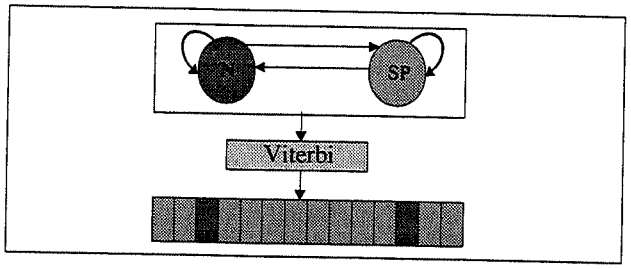


Figure 1: *Processus itératif : initialisation*

Dans la partie itérative du processus (figure 2), les différentes phases sont :

1. A partir des zones indexées Parole, le modèle le plus probable est détecté parmi les modèles de locuteurs qui n'ont pas encore été introduits dans le modèle de conversation. Cette phase est réalisée à l'aide de l'algorithme SWGM[2] [8].
2. Un nouvel état représentant le locuteur choisi en (1) est ajouté au modèle de Markov. Les probabilités d'émission de cet état sont calculées. Les poids des transitions sont adaptés pour prendre en considération le nouveau nombre d'états.
3. L'algorithme de Viterbi est appliqué pour obtenir l'alignement optimal par rapport à la topologie actuelle du modèle de Markov. On obtient une indexation temporaire, qui sera de nouveau remise en cause à l'étape suivante.
4. A la fin de l'itération, le critère d'arrêt est testé : l'indexation actuelle est-elle meilleure que l'indexation précédente ? Si un gain est constaté, une nouvelle itération commence.

NB : En pratique, le système teste un deuxième critère d'arrêt : Reste-t-il des blocs étiquetés Parole pour la sélection et l'ajout d'un nouveau modèle de locuteur ?

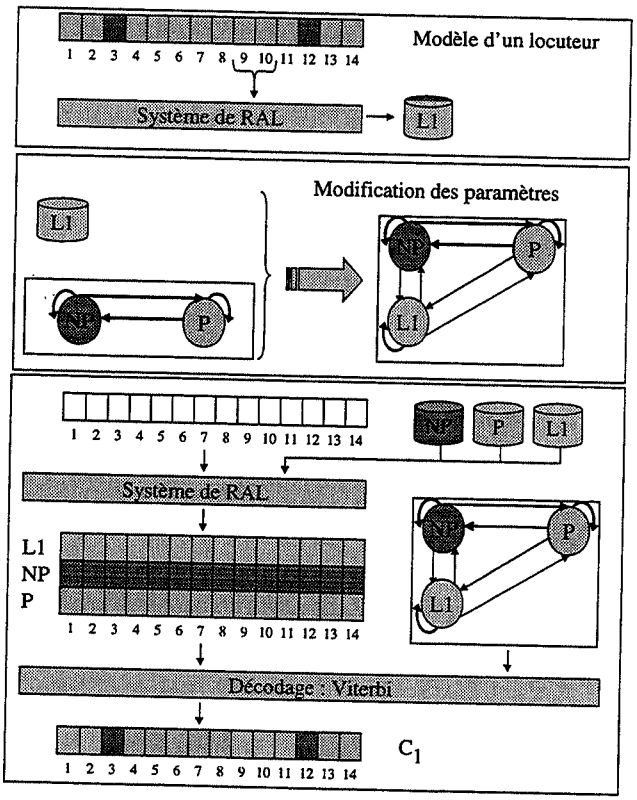


Figure 2: *Processus itératif : traitement du premier locuteur détecté*

3. Système de vérification du locuteur

Les modèles de locuteurs employés et l'ensemble des probabilités d'émission sont calculés par le système de reconnaissance du locuteur AMIRAL[3] [5], développé au LIA.

La paramétrisation acoustique (coefficients cepstraux) est effectuée grâce au module développé par le consortium ELISA[4].

[2]Sorted Weighted Geometrical Mean

[3]Architecture Multi-reconnaisseurs pour l'Indexation et la Reconnaissance Automatique du Locuteur

[4]Le consortium ELISA est composé de laboratoires de

Les locuteurs sont modélisés par des mixtures de gaussiennes (GMM à 16 composantes avec une matrice de covariance pleine [7]) apprises par l'algorithme EM[5] [6] (critère Maximum Likelihood).

Une spécificité d'AMIRAL est de considérer le signal de parole au niveau de blocs de 0,3 s (= 30 trames), sur lesquels des scores de vraisemblance normalisés (par modèle du monde et MAP[6] [4]) sont calculés. La normalisation MAP employée permet d'obtenir des scores assimilables à des probabilités. AMIRAL prend également en charge la sélection des modèles à l'aide de l'algorithme SWGM, adapté à l'identification du locuteur à partir d'enregistrement pluri-locuteurs.

4. Expériences

Les expériences sont réalisées pour vérifier la validité :

- du modèle de conversation proposé (en particulier les changements de locuteur),
- de l'algorithme itératif (détection d'un nouveau locuteur, arrêt de l'ajout de modèle).

4.1. Ensembles de données

La méthode proposée dans cet article a été expérimentée sur un sous-ensemble de données issues de la campagne d'évaluation NIST 1998. Les messages ont été simulés à partir de la concaténation de segments de parole téléphonique (conversation réelle) mono-locuteur issus de la base Switchboard.

Deux sous-ensembles indépendants (définis par le consortium ELISA) sont utilisés :

- Un premier ensemble de 25 locuteurs est consacré au développement (Dev).
- Un ensemble Eva permet la validation des paramètres mis au point sur Dev. Ce corpus a la même taille et la même structure que Dev. Les populations de locuteurs de Dev et Eva sont disjointes.

Pour chaque sous-ensemble, nous disposons de 2 minutes de parole par locuteur pour l'apprentissage du modèle et de 30 secondes pour la génération des messages de test.

4.2. Génération des messages

Les messages sont générés par concaténation de blocs (0,3s) de parole mono-locuteur. La méthode utilisée est :

- l différents locuteurs sont sélectionnés parmi les 25 locuteurs disponibles.
- i (*avec* $i \geq l$) différents segments sont choisis, tels que chaque locuteur soit présent au moins une fois.
- La durée d de chaque intervention est déterminée.

l, i and d sont des nombres tirés aléatoirement dans des distributions gaussiennes (Table 1). La sélection des locuteurs, l'ordre d'apparition des segments et le choix des segments sont générés à partir de distributions uniformes.

5000 messages ont été générés pour chaque corpus.

NB : Il n'existe pas de situation où deux locuteurs parlent en même temps.

Paramètres	Moyenne	écart type
l	5	1
i	15	2
d (# 0,3s)	10	2

Table 1: *Paramètres pour la génération des données*

4.3. Tests

Les critères sélectionnés pour mesurer la performance du système sont :
• Le pourcentage de blocs correctement indexés (BC). Ce pourcentage correspond au taux de blocs pour lesquels le système propose une étiquette conforme à l'identité réelle.
• Le pourcentage de blocs mal indexés (BM). Il est calculé sur les blocs pour lesquels le système ne propose pas la bonne identité.

Un taux d'erreur d'attribution d'un bloc à un état (EA) est calculé à partir de ces deux valeurs :

$$EA = \frac{BM}{BM + BC} \quad (1)$$

EA permet de mesurer la justesse de l'indexation proposée.

Deux valeurs mesurent le taux d'indécision du système :
• Le pourcentage de segments étiquetés Non Parole (BNP).
• Le pourcentage de segments étiquetés Parole (BP).

La somme de ces taux indique le pourcentage de blocs pour lesquels le système ne prend pas de décision.

NB : la segmentation de réference, issue de la génération des messages, n'attribue pas de bloc pour les étiquettes NP et P. Les blocs affectés à NP et P ne sont donc pas comptabilisées lors du calcul de EA.

4.4. Résultats

Les résultats portés dans Table 2 montrent un taux global de 66% de blocs correctement indexés sur Dev (respectivement 62,1% sur Eva).

Le taux d'erreur d'attribution (EA) reste élevé (environ 30% sur Dev et Eva) ; mais le taux d'erreur obtenu

recherche européens qui travaillent sur une plate-forme commune. Les laboratoires d'ELISA ayant participé à NIST 1999 sont : ENST (France), EPFL (Suisse), IDIAP (Suisse), IRISA (France), LIA (France), RIMO — Rice (USA) et Mons (Belgique) —, RMA (Belgique), VUTBR (République Tchèque).

[5]Expectation-Maximization

[6]Maximum A Posteriori

Corpus	Décisions			EA
	BC	BM	Total	
Dev	66,0%	28,6%	94,6%	30,2%
Eva	62,1%	26,4%	88,5%	29,8%

Table 2: *Décisions du système de suivi de locuteurs : Taux calculés sur Dev et Eva (5000 messages chacun). Pour tous les blocs : BC= % de blocs correctement indexés, BM= % de blocs mal indexés, EA= taux d'erreur d'attribution.*

par une simple décision bayésienne[7] (équ. 2) atteint 54,5% d'erreur d'attribution sur Dev (et 53,8% sur Eva). Ce dernier résultat montre que le taux d'erreur d'affectation provient principalement de la difficulté intrinseque de la base Switchboard et des erreurs du système AMIRAL.

$$dec_{Bay}(i) = \underset{l \in M}{\text{ArgMax}}(p(s_i|M_l)) \qquad (2)$$

NB : M est l'ensemble des modèles de locuteurs. s_i est le $i^{ième}$ bloc du signal.

Le taux d'indécision observé (blocs attribués à aucun modèle de locuteur) est de 5,4% (BP + BNP) sur Dev, respectivement 11,5% sur EVA (Table 3). Le taux obtenu est faible au vu de la difficulté de la tâche.

Corpus	Indécisions		
	BNP	BP	Total
Dev	0,9%	4,5%	5,4%
Eva	1,6%	9,9%	11,5%

Table 3: *Indécisions du système de suivi de locuteurs : Taux calculés sur Dev et Eva (5000 messages chacun). Pour tous les blocs : BP= % blocs attribués à Parole, BNP= % blocs attribués à Non Parole.*

Le taux de locuteurs détectés dans les messages est proche de 90% sur Dev et Eva. Néanmoins, le nombre de locuteurs ajoutés à tort au modèle de Markov est important : environ 70% des locuteurs ajoutés sont des locuteurs qui ne font pas partie du message, que ce soit sur Dev ou Eva. Ces locuteurs représentent environ les 2/3 des erreurs d'attribution relevées durant les tests (EA $\simeq$ 30% sur Dev et Eva).

5. Conclusion

Dans cet article, le système de suivi de locuteurs utilise un modèle de Markov évolutif pour modéliser la conversation et pour déterminer automatiquement les locuteurs présents dans les messages. L'approche est basée sur un algorithme itératif qui détecte et ajoute les modèles de locuteur un à un. A chaque étape, une indexation est proposée, fonction de l'ensemble des connaissances disponibles. Cette indexation est remise en cause à l'itération suivante jusqu'à trouver la solution optimale (détection de tous les locuteurs du message).

[7] pour un bloc donné, le locuteur le plus probable est choisi parmi les 25 locuteurs du corpus

Les résultats obtenus sont encourageants au vu du taux d'erreur d'attribution et du taux de non décision. Cependant, trop de locuteurs absents des messages sont ajoutés au modèle de conversation.

Les travaux futurs devront porter sur trois points :

- La méthode de sélection des locuteurs à ajouter doit être améliorée.
- Actuellement, le modèle de Markov n'utilise pas de modèle de durée explicite, l'utilisation d'un tel modèle devrait permettre de limiter les problèmes de sur-segmentation.
- Le système actuel est adapté aux tâches de suivi de locuteurs, il sera étendu à des tâches d'indexation, où les modèles de locuteurs doivent être construits, ou adaptés, à partir des données du message.

Bibliographie

[1] P. Delacourt, D. Kryze, C.J. Wellekens. Use of second order statistic for speaker-based segmentation, *EUROSPEECH*, 1999.

[2] H. Gish, H-H Siu, R. Rohlicek. Segregation of speakers for speech recognition and speaker identification, *ICASSP*, pages 873-876, 1991.

[3] K. Sönmez, L. Heck, M. Weintraub, Speaker tracking and detection with multiple speakers, *EUROSPEECH*, 1999.

[4] T. Matsui, S. Furui. Likelihood normalization for speaker verification using a phoneme and speaker-independent model, *Speech communication*, pages 109-116, August 1995.

[5] C. Fredouille, J.-F. Bonastre, T. Merlin, Similarity normalization method based on world model and a posteriori probability for speaker verification, *EUROSPEECH*, 1999.

[6] D. Dempster, N. Larid, D. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.*, Vol. 39, pp 1-38, 1977.

[7] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models, *Speech Communication*, pp 91-108, Aug. 1995.

[8] J-F. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, C. Wellekens, Différentes stratégies pour le suivi de locuteur, *RFIA 2000*, Jan. 2000.

Phonétique/Phonologie

Les *euh* et les allongements dits « d'hésitation » : deux phénomènes soumis à certaines contraintes en français oral non lu

Maria CANDEA

Institut de Phonétique, Université de Paris III, 19, rue des Bernardins 75005 – Paris, France
Mél: candea@ext.jussieu.fr

ABSTRACT

This article studies two phenomena abusively called «hesitant», in spontaneous French speech : «euh» and vowel lengthening. Our hypothesis is that these two phenomena are in complementary distribution : vowel lengthening affects almost exclusively syllables of the type (C)V of empty words whereas «euh» is much more often found following full words or (C)VC type syllables of empty words. It is interesting to note that connectives behave both as empty words (they combine with vowel lengthening if their final syllable is of the type (C)V), and as full words (they combine with «euh») ; we attribute this behavior to their enunciative role and to their intonation contour (generally high and/or modulated) which is markedly different from that of other empty words.

1. INTRODUCTION

A la différence des pauses silencieuses dans la parole qui ont fait l'objet de nombreuses études et qui bénéficient en outre d'une longue tradition issue des précis et manuels de déclamation théâtrale, les marques dites « d'hésitation » n'ont commencé à être étudiées que très tardivement, vers la fin des années 50 et uniquement en anglais, grâce à l'ampleur que commençait à prendre le recours à des corpus de parole naturelle. Goldman-Eisler —surtout [Gol68]—, d'une part, et Maclay et Osgood [Mac59], d'autre part, ont reconnu à cette époque, dans ces phénomènes, des objets d'étude pour la linguistique et ont publié des travaux qui sont réellement, à en juger par toutes les bibliographies postérieures, à la base de tous les travaux ultérieurs portant sur ce sujet.

La toute première étude sur le français prenant en compte ces phénomènes (Grosjean & Deschamps, [Gro72] paraît seulement en 1972 et sera suivie de deux autres études en 1973 et 1975, faisant systématiquement référence aux résultats et aux travaux de [Gol68].

Les marques dites « d'hésitation » qui avaient à l'origine intéressé les psycholinguistes anglo-saxons et qui ont ensuite intéressé d'autres chercheurs (voir à ce propos Duez [Due91] qui passe en revue les principales études anglophones entre 1958 et 1987) sont actuellement systématiquement prises en compte dans les systèmes de reconnaissance automatique de la parole spontanée. Cette prise en compte est la plupart du temps purement empirique, les essais de théorisation sont à leurs débuts (voir par ex. [Bea92] pour l'anglais et les travaux de Guaïtella qui est à notre connaissance le seul auteur ayant clairement essayé de décrire cet aspect pour le français).

En ce qui nous concerne, nos recherches portent sur les caractéristiques et la distribution de ces marques (euh — ex : *il achetait euh des fusains* ; allongements vocaliques — ex : *c'était à : : : Villiers* et répétitions de mots outils grammaticaux — ex : *celui du bébé ours qui euh qui lui va à merveille*). Nous étudions également les nombreuses possibilités de combinaison de ces marques entre elles et avec la pause silencieuse, combinaisons fréquentes en français oral non lu (dans nos corpus tout comme dans ceux qui ont été étudiés par d'autres chercheurs, [Gro72], [Due91], [Gua91]... Notre corpus actif est constitué de 70 minutes de parole (env. 10.000 mots, 11 locuteurs, âgés de 13-14 ans enregistrés en classe de français) et nous avons entrepris actuellement de vérifier nos résultats les plus significatifs à partir d'une trentaine de minutes extraites de corpus très variés enregistrés par notre équipe de recherches de l'Univ. de Paris III.

2. DISCUSSION SUR LA TERMINOLOGIE

La terminologie utilisée dans les rares études systématiques de ces marques en français est extrêmement hétérogène (les auteurs parlent de phénomènes d'hésitation, pauses sonores, pauses non silencieuses, pauses remplies, pauses pleines...). En revanche, en français ou en anglais, le mot *hésitation* revient régulièrement pour désigner ces marques ou le processus cognitif qu'elles sont censées indiquer.

Le choix de ce mot n'est pratiquement jamais justifié et provient à notre avis, plus ou moins directement de l'étude de [Mac59] qui a été la première à trancher explicitement en faveur de l'emploi du terme générique « hesitation phenomena » au détriment des termes de type « disturbances » ou « disfluencies » qui étaient à l'époque en concurrence.

L'originalité de [Mac59] et son énorme impact par la suite ont fait que le choix de ce terme n'a, à notre connaissance, jamais été contesté depuis, alors qu'aucune étude scientifique n'a pu mettre en évidence un rapport systématique et obligatoire de cause à effet entre ces marques et un processus cognitif

d'« hésitation » dans le sens courant donné par le Petit Robert (1996) de « être dans un état d'incertitude, d'irrésolution qui suspend l'action, la détermination », sens qui implique surtout la difficulté de choisir entre deux ou plusieurs possibilités.

Or, les différentes approches cognitivistes de ce type de phénomènes ont surtout mis en évidence un rapport entre ces marques et l'effort d'encodage du locuteur. Ces marques signalent une difficulté due à un simple retard dans la « programmation des unités » ou bien à une difficulté passagère de « conceptualisation des unités » [Due91], autrement dit elles représentent une « activité métacognitive » dirigée vers l'auditeur qui accompagne l'activité cognitive de recherche/ production d'une unité linguistique par le locuteur et que l'auditeur serait capable de décoder en tant qu'indice métacognitif [Bre95].

Les chercheurs s'accordent pour dire que la durée est le paramètre le plus saillant de ces marques (même si elle ne suffit pas toujours pour les définir et les reconnaître) ; cette durée n'est souvent pas due à un « embarras du choix » de la part du locuteur, n'est pas un indice « d'irrésolution » mais tout simplement un temps d'encodage plus long que prévu et qui nécessite un fort ralentissement ponctuel du rythme.

La durée n'est toutefois pas caractéristique pour les répétitions de mots outils ont un fonctionnement différent des *euh* et des allongements vocaliques : en effet, dans tous les corpus que nous avons pu étudier, la durée moyenne qu'on relève entre le début de la répétition et le début du mot cible est significativement inférieure à la durée moyenne qu'on relève entre le début du *euh* ou de l'allongement vocalique et le début du mot cible (test Mann-Whitney, $p<0,001$), cette différence étant surtout due à la durée de la pause silencieuse qui suit immédiatement chacune de ces marques et non pas à leur durée intrinsèque.

Ainsi, le processus d'« hésitation » (irrésolution devant plusieurs choix en concurrence) ne peut pas être associé systématiquement à la production de ces marques par un locuteur donné. Le terme « hésitation » qui est en train de s'imposer dans la littérature francophone à partir de la littérature anglophone ne nous semble par conséquent pas adéquat, même s'il est sans doute mieux choisi que les termes « disturbances » et « disfluencies » que Maclay et Osgood ont voulu éviter, car ces termes étaient trop cliniques et tendaient à ranger ce type de phénomènes du côté des pathologies du langage.

En ce qui nous concerne nous avions adopté dans un premier temps la proposition plus neutre de Grosjean et Deschamps [Gro72] qui parlaient de « pauses sonores » (voir aussi [Due91], [Can97]), mais après avoir approfondi l'étude d'une grande quantité de corpus nous pensons que ce terme n'est à son tour pas suffisamment neutre. En effet, si la nature acoustique des pauses silencieuses est constante celle des 'pauses sonores' ne l'est pas, (on y rencontre toutes les voyelles et même certaines consonnes). Les appeler 'pauses' implique une indifférenciation théorique qui n'est pas établie (au contraire, cf. [Swe98] les 'pauses sonores' *uh* et *um* en anglais n'auraient pas le même rôle...) La proposition terminologique qui nous a semblé la plus neutre est celle de [Mor98]. Les auteurs regroupent ces phénomènes sous le nom de « marques du travail de formulation », dans une optique d'analyse énonciative. C'est le choix que nous faisons également par la suite (abr : marques du TdF).

3. *EUH* /VS/ ALLONGEMENT VOCALIQUE ?

Si les répétitions de mots outils peuvent être isolées des deux autres marques du TdF notamment en raison de leur durée, il n'en est rien en ce qui concerne les *euh* et les allongements vocaliques finals. Les rares chercheurs qui ont pris en compte ces phénomènes dans leurs études ont des avis divergeants en ce qui concerne le fonctionnement de ces deux marques.

En effet, dans les premières études sur le français, principalement [Gro75], Grosjean et Deschamps classent les *euh* du côté du temps total d'élocution, de même que les allongements vocaliques. Les auteurs relèvent des pourcentages différents pour l'anglais et le français : l'anglais privilégierait nettement les *fillers* de type *uh/um* par rapport aux allongements alors que le français aurait seulement une légère préférence, moins marquée, pour les *euh*. Les auteurs attribuent cette différence principalement aux structures syllabiques prédominantes dans les deux langues (syllabes ouvertes en français, syllabes fermées en anglais), et non pas aux différences idiolectales entre les locuteurs. Ils considèrent que les deux marques auraient le même rôle et le même fonctionnement et que leurs pourcentages cumulés seraient stables (plus il y a d'allongements moins il y a de *euh* et vice versa).

Duez conteste partiellement ce point de vue et propose de regrouper les *euh* du côté du temps total de pause et de laisser uniquement les allongements vocaliques du côté du temps total d'élocution. Ce regroupement évite notamment de considérer chaque *euh* comme étant une syllabe et de fausser ainsi, en raison de la longueur exceptionnelle de nombreux *euh*, la durée moyenne des syllabes. Néanmoins, il ne ressort pas clairement de son ouvrage qu'elle attribuerait des rôles différents aux allongements vocaliques et aux *euh* : en effet, lorsqu'elle présente brièvement la distribution des « pauses sonores » Duez regroupe les deux types de marques et signale leur combinatoire très similaire avec la pause silencieuse ([Due91], pp.71-78).

A la même époque, dans [Gua91] Guaïtella décide de confondre complètement les deux marques sous le

nom de « hésitations vocales » (elle ne fait aucune distinction dans ses comptages entre ces deux types de marques, mais crée une catégorie à part pour les *euh* brefs), sans pour autant donner une justification théorique à ce choix.

Plus récemment, dans une approche (co)énonciative de la prosodie, les auteurs de [Mor98] pensent que les allongements vocaliques finals n'ont pas la même distribution syntaxique que les *euh* et avancent l'hypothèse que ces deux marques du TdF pourraient avoir des 'portées' et des rôles différents (les allongements porteraient sur une séquence cible plus limitée que les séquences introduites par un *euh*). Cette hypothèse est encore à l'étude et n'est, pour le moment, pas validée statistiquement.

Afin d'y voir plus clair devant ces points de vue aussi divergeants, nous avons tâché d'analyser plus en détail le contexte immédiat de ces deux marques et tenté de dégager d'éventuelles contraintes combinatoires.

3.1 Contrainte lexicale

En étiquetant les catégories d'unités qui portaient dans notre corpus un allongement vocalique marque du TdF (allongement à contour mélodique bas et ayant une durée significative, c'est-à-dire supérieure à celle d'une syllabe accentuée après application des facteurs de pondération de la durée intrinsèque de la voyelle) nous nous sommes aperçue que le nombre de mots outils (abr : MO) était bien supérieur à celui des mots pleins (abr :MP) : 258 MO portant un tel allongement contre 26 MP, soit 90,85% MO contre 9,15% MP). Nous avons ainsi constaté une très nette préférence pour l'allongement des MO et une forte tendance à éviter l'allongement des MP.

Ce résultat est en outre concordant avec celui obtenu par [Gro72] et [Gro73] (88,75% et 94,16% des allongements portant sur des MO, en fonction du corpus) ; nous ne connaissons malheureusement pas d'autre étude qui ait fait ce type de décompte sur d'autres corpus.

Même s'il est encore prématuré de l'affirmer catégoriquement, ces résultats concordants obtenus à partir de corpus très différents nous permettent de formuler l'hypothèse selon laquelle le français oral non lu aurait largement tendance à faire porter les allongements vocaliques marques du TdF sur des MO et non sur des MP.

Cette remarque ne suffirait cependant pas pour isoler la distribution du *euh* par rapport à celle des allongements : il faudrait pour cela appliquer le même critère pour la distribution des *euh*. Nous n'avons pas trouvé de données en ce sens pour le français, ([Gro72] et [Gro73] ont appliqué le critère mot plein/mot outil uniquement aux allongements.)

Nous avons appliqué ce critère sur notre corpus actif, (après avoir éliminé les *euh* placés en tout début de prise de parole et les *euh* précédés par une pause silencieuse longue supérieure à 2 secondes et après avoir également éliminé provisoirement les *euh* précédés d'un connecteur, voir *infra*) nous avons obtenu, sur les 328 occurrences de *euh* restantes, un pourcentage de 17,07% de *euh* précédés d'un MO et un pourcentage de 82,93% de *euh* précédés d'un MP. Ce pourcentage de 17,07% est déjà très significativement différent de celui obtenu pour les allongements (91,25% des allongements précédés par un MO, écart-type 2,73). L'écart entre ces deux pourcentages sera encore plus fort lorsque nous isolerons les MO à structure syllabique (C)VC.

3.2 Contrainte syllabique

En effet, en nous inspirant de l'hypothèse formulée *a priori* dans [Gro75] mais non démontrée, selon laquelle la structure syllabique ouverte /vs/ fermée aurait une influence sur la fréquence des allongements vocaliques marques du TdF, nous avons voulu vérifier ce qu'il en était à partir des données de notre corpus.

En observant la structure syllabique des 26 MP qui portaient un tel allongement sur la syllabe finale, nous avons relevé un seul exemple de mot finissant par une syllabe de type (C)VC (après avoir appliqué les facteurs de pondération de la durée, sachant qu'une voyelle appartenant à une syllabe fermée a une durée intrinsèque supérieure à celle appartenant à une syllabe ouverte). Il est vrai que le nombre d'exemples issus de notre corpus n'est pas suffisant pour savoir si ce résultat est significatif ou non.

D'autre part, en observant la structure syllabique des MO porteurs d'un allongement de ce type, nous avons relevé seulement 3 occurrences de MO à syllabe fermée (il s'agit de trois monosyllabiques, *elle*, *donc*, et *une*) soit 1,16% des 258 exemples, ce qui est en revanche très significatif.

Ces résultats mettent en évidence une forte tendance en français oral non lu à éviter les allongements de syllabes fermées ; cependant ces résultats ne suffiraient pas en eux-mêmes pour prévoir le comportement des locuteurs dans les cas où ils seraient amenés à marquer le TdF sur une syllabe fermée.

Or, en revenant aux données obtenues pour les contextes de type « MO suivi de *euh* » (17,07% des contextes) nous nous sommes aperçue que largement plus de la moitié, 34 occurrences, soit 10,37% du total des contextes avant *euh* étaient des MO de type (C)VC (principalement *elle, sur, avec, une*), et seulement 22 occurrences, soit 6,70% du total des contextes avant *euh* étaient des MO de type (C)V. (voir table 1).

Table 1 : Distribution des allongements et des *euh* en fonction du contexte mot outil (MO) ou mot plein (MP) et en fonction de la structure syllabique des MO

	allongement	*euh*
MP allongé /vs/ suivi de euh	26	**272**
MO allongé /vs/ suivi de euh	**258**	56
MO (C)V allongé /vs/ suivi de euh	**255**	22
MO (C)VC allongé /vs/ suivi de euh	3	**34**

Ces résultats montrent que, dans notre corpus, les MO ont tendance à être largement plus souvent marqués par un allongement indiquant le TdF plutôt que d'être suivis d'un *euh*, avec néanmoins une restriction portant sur la structure de la syllabe allongée : lorsque la syllabe allongée est de type (C)VC il est beaucoup plus fréquent que ces MO soient suivis d'un *euh*.

Cette contrainte syllabique ne semble pas jouer sur les MP qui sont de toute manière très rarement porteurs de ce type d'allongement ; il n'en reste pas moins, que dans notre corpus, les 26 mots pleins allongés finissent, à une seule exception près, par une syllabe de type CV.

3.3 Le cas des connecteurs

Dans notre corpus la classe des connecteurs (conjonctions et adverbes introducteurs, les exemples les plus fréquents étant *et, alors, mais, donc, puis, et puis, et alors, ben, après, si*) a un comportement combinatoire différent du reste du corpus par rapport aux deux marques du TdF qui nous intéressent, le *euh* et les allongements (ce comportement différent semble être confirmé par les travaux de Morel et Danon-Boileau à partir d'une grande variété de corpus, [Mor98]). Ces unités se combinent en effet très fréquemment avec le *euh* (169 exemples dans notre corpus) quelle que soit leur structure syllabique. Si la structure syllabique est ouverte, ils peuvent se combiner aussi avec l'allongement.

Ce comportement est en fait identique à celui des MP à ceci près que la combinaison avec l'allongement est plus fréquente est s'approche de celle que l'on constate pour les MO. Nous pensons que l'explication est à chercher dans le rôle énonciatif de ces unités à l'oral spontané : en effet, ces connecteurs présentent bien souvent un contour mélodique très montant ou modulé en cloche ce qui est extrêmement rare pour les autres types de mots outils (voir aussi [Mor98]). Selon nous, la présence de ce contour interdit en français l'allongement de la voyelle qui le porte par un palier plat sans modification spectrale obligatoire de cette voyelle (elle se « transforme » en fait en *euh* dès le début du contour plat typique du TdF).

4. CONCLUSIONS

Les données que nous avons assemblées à partir de notre corpus nous incitent à formuler l'hypothèse que les deux marques du TdF étudiées, le *euh* et les allongements, auraient en français oral non lu une distribution complémentaire et seraient pour ainsi dire des variantes combinatoires d'une seule et même marque. Cette hypothèse n'est pas en contradiction avec la stabilité des pourcentages cumulés de ces deux marques évoquée dans [Gro75] et n'est pas non plus en contradiction avec l'hypothèse de la 'portée' différente de ces deux marques de [Mor98]. Si nos hypothèses sont validées sur un plus grand nombre d'enregistrements, nous pensons que cette portée différente s'expliquerait par la distribution syntaxique différente des MO et des MP en français et non pas par une spécialisation énonciative de chacune de ces deux marques (il s'agirait d'une simple corrélation contextuelle et non d'une relation de cause à effet).

BIBLIOGRAPHIE

[Bea92] Bear J. et *alii* (1992) "Detection and correction of repairs in human-computer dialog", Proc. of the Annual Meeting of the Association for Computational Linguistics, Delaware

[Bre95] Brennan, S.E., Williams, M, (1995) "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about metacognitive states of speakers", Journal of Memory and Language, 34, pp. 383-398

[Can97] Candea M. (1997) "Peut-on définir la pause dans le discours comme un lieu d'absence de toute marque?", Travaux linguistiques du CERLICO, 10, pp. 231-244

[Due91] Duez D. (1991), La pause dans la parole de l'homme politique, CNRS

[Gol68] Goldman-Eisler F. (1968) Psycholinguistics : experiments in spontaneous speech, Academic P

[Gro72] Grosjean F., Deschamps A. (1972-73) "Analyse
[Gro73] des variables temporelles du français spontané",
Phonetica, 26, 130-156, 28, pp.191-226 et 31,
[Gro75] pp.143-183

[Gua91] Guaïtella, I. (1991) "Hésitations vocales en parole spontanée : réalisations acoustiques et fonctions rythmiques", Travaux de l'Institut de Phonétique d'Aix, vol.14, pp. 113-130

[Mac59] Maclay H., Osgood Ch.E. (1959) "Hesitation Phenomena in Spontaneous English Speech", Word, 15 (4), pp. 19-44

[Mor98] Morel M., Danon-Boileau L. (1998) Grammaire de l'intonation. L'exemple du français, Ophrys

[Swe98] Swerts, M. (1998) "Filled pauses as markers of discourse structure", Journal of Pragmatics, 30 (4), pp. 485-496

Étude sur l'implémentation du schwa pour quatre locuteurs berbères de tachelhit

Naïma Louali & Gilbert Puech

UMR Dynamique du Langage
ISH 14 avenue Berthelot 69363 Lyon Cedex 07
Tel +33 (0)4 72 72 64 93 / Fax : +33 (0)04 72 72 65 90
e-mail : nlouali@ish-lyon.cnrs.fr/ puech@univ-lyon2.fr

ABSTRACT

In Tashlhiyt, a Berber dialect spoken in southern Morocco, consonants may occupy the position of syllabic nucleus. According to Dell [96], a short vocoid, referred to here as schwa, may occur when a syllabic consonant is voiced but is ruled out when it is unvoiced. This paper investigates the latter context for four speakers of Tashlhiyt with respect to the occurrence of schwa. Dell & Elmedlaoui's hypothesis is to some extent consistent with the data analyzed for one speaker but not for the three others. The characteristics of schwa for duration and place in the vocalic space are analyzed for speakers producing schwa. A phonological interpretation of the whole set of data is eventually sketched out, in accordance with Cole [96].

1. INTRODUCTION

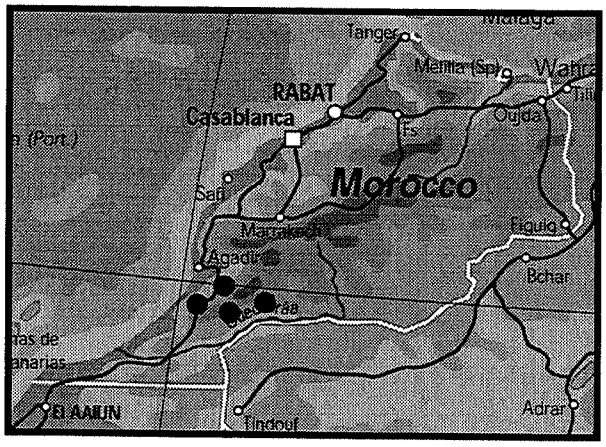
Le domaine berbère inclut plusieurs ensembles dialectaux, dont le tachelhit parlé dans le Haut et Anti-Atlas (sud du Maroc). Ce dialecte a récemment suscité l'engouement des phonologues pour la capacité d'une de ses variétés, celle parlée par Elmedlaoui originaire d'Imdlawn, à faire jouer à toutes les consonnes le rôle de centre de syllabe, y compris dans des contextes où la présence phonétique d'un vocoïde, appelé communément schwa, est en général avérée dans d'autres variétés. [Lou96], [Cole99].

Dell et Elmedlaoui affirment que la présence d'un schwa est exclue pour certaines formes en "Imdlawn Tashlhiyt Berber (ITB) : "As pointed out in 1985, in ITB schwas are always adjacent to a voiced consonant. This generalization is true of all styles of elocution, even the most deliberate ones. Take for instance the phrase /t-s-qssf-t≠stt/ 'you shrank it'. [...] It contains only voiceless consonants in its underlying representation, and it must be pronounced voiceless from one hand to the other : [tsqssftstt]. Voiced vocoïd cannot be inserted in it at any point [...]. " [Dell 96, p. 225].

La présente étude porte sur l'implémentation du schwa, pour quatre locuteurs tachelhit dans le contexte de consonnes non voisées. Elle met en évidence que :

1) les locuteurs ont des comportements différents ;
2) la probabilité de l'occurrence du schwa augmente avec la complexité consonantique de la forme.

On trouve certes des réalisations formées d'une séquence de consonnes non voisées mais ces réalisations peuvent alter-ner chez le même locuteur avec une séquence comportant un schwa. L'ITB constitue un cas à part dans la mesure où son locuteur n'accepte pas la grammaticalité de certaines formes avec un schwa. Cet idiome est particulièrement intéressant parce qu'il constitue un cas unique dans la typologie comme le montre Zec [Zec 95] et qu'il a servi de support à différentes modélisations des données comme celle de Clements [Cle 97].



Localisation des quatre points d'enquête

2. PRÉSENCE PHONÉTIQUE DU SCHWA

Un corpus a été enregistré avec quatre locuteurs tachelhit originaires des localités suivantes: Tiznit, Tanalt, Anezi, Massa (cf. carte). Ce corpus comprend les items suivants :

/kf/	"donne !"
/kf≠t/	"donne-le !"
/ks/	"pais !"
/ks≠t/	"pais-le (= fais-le paître !)"
/t+kʃf/	"elle est décolorée"

Ces formes étaient transcrites en caractères arabes sans vocalisation. Chaque forme est pointée avec en accompagnement la question : *magg ghwa* ? (qu'est ce que c'est ?). On a ainsi obtenu dix répétitions non consécutives pour chacune des formes. Toutefois nous n'avons retenu que neuf pour l'analyse éliminant ainsi les échantillons correspondant à la première répétition. Par ailleurs la forme /tkʃf/ est un emprunt arabe qui bien qu'intégré à la langue a été refusé par deux locuteurs.

L'examen acoustique de ces productions se répartit en trois types de cas, on constate :

1) un vocoïde d'une durée supérieure à 20 ms comme par exemple dans la figure 1 : le schwa est alors transcrit comme une voyelle brève, au même titre que les voyelles périphériques i, u, a ;

2) un vocoïde formé de trois périodes au moins mais d'une durée inférieure à 20 ms : nous utilisons un schwa suscrit pour ce vocoïde transitionnel (exemple en figure 2) ;

3) l'absence de vocoïde (figure 3).

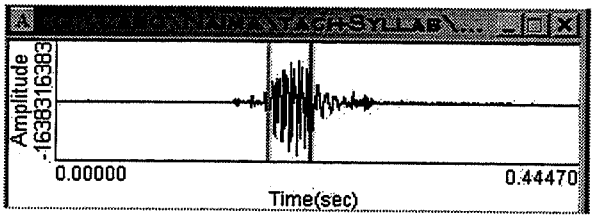


Figure 1. Signal pour une production de [kəf]

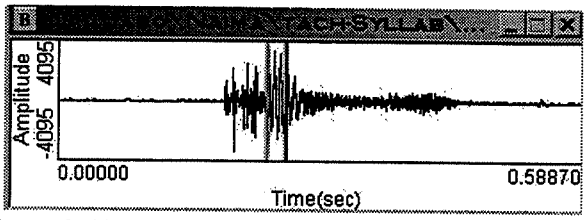


Figure 2. Signal pour une production de [k°f]

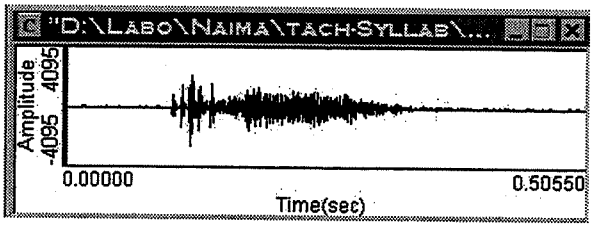


Figure 3. Signal pour une production de [kf]

Nous examinerons successivement le cas de chaque locuteur.

Le locuteur de Tiznit produit pour la grande majorité des items (38 fois sur 45) un vocoïde plein, dans deux cas seulement un vocoïde transitionnel et dans cinq cas une forme sans trace de vocoïde (tableau 1). Le F0 sur les schwas mesurés est en moyenne de 150 Hz.

Le locuteur originaire de Tanalt ne produit pas de vocoïde pour les formes biconsonantiques ; le schwa est présent une fois sur deux en moyenne pour les formes triconsonantiques et de façon presque systématique pour la forme /tkʃf/ (tableau 2). Le F0 sur les schwas mesurés est en moyenne de 178 Hz.

Le troisième locuteur originaire de Anezi produit six vocoïdes pleins, sept vocoïdes transitionnels et vingt-deux formes sans vocoïde (tableau 3). Le F0 sur les schwas mesurés est en moyenne de 155 Hz.

Le quatrième locuteur originaire de Massa produit les quatre formes qu'il accepte (l'emprunt arabe étant refusé) sans vocoïde.

ks	0	61	46	0	61	53	53	0	45
kf	34	61	48	15	16	41	42	0	51
kst	0	58	48	46	60	52	40	48	49
kft	44	46	39	39	58	62	50	41	53
tkSf	65	56	49	52	50	54	52	74	41

Tableau 1. Occurrence et durée en ms du schwa pour le locuteur de Tiznit.

ks	0	0	0	0	0	0	0	0	0
kf	0	0	0	0	0	0	0	0	0
kst	0	0	0	0	33	31	31	0	0
kft	0	0	0	0	41	40	43	0	32
tkSf	0	31	16	58	46	43	51	12	46

Tableau 2. Occurrence et durée en ms du schwa pour le locuteur de Tanalt.

ks	0	0	0	0	34	0	19	17	0
kf	28	25	0	17	0	0	0	0	0
kst	36	22	0	21	0	0	0	13	0
kft	36	15	0	19	16	0	0	0	0

Tableau 3. Occurrence et durée en ms du schwa pour le locuteur de Anezi.

ks	0	0	0	0	0	0	0	0	0
kf	0	0	0	0	0	0	0	0	0
kst	0	0	0	0	0	0	0	0	0
kft	0	0	0	0	0	0	0	0	0

Tableau 4. Cas du locuteur de Massa

La figure 4 ci-dessous met en évidence les deux paramètres de variabilité qui ressortent des tableaux 1 à 4 :

1) variabilité entre locuteurs,

2) variabilité due à la complexité consonantique de la séquence : la fréquence d'occurrence du schwa augmente avec le nombre de consonnes composant la forme.

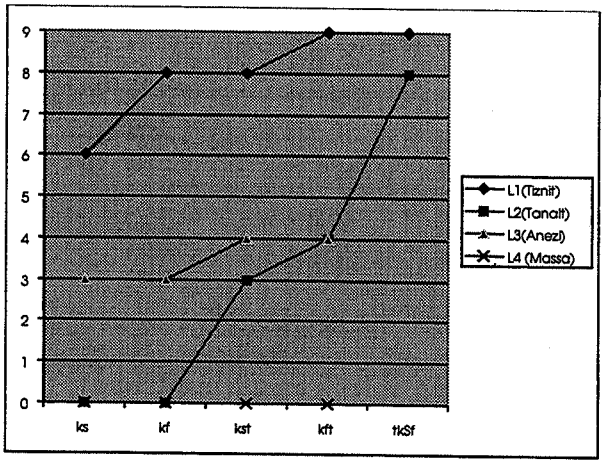


Figure 4. Variabilité d'occurrence du schwa

3. DURÉE DU SCHWA

Nous avons comparé la durée des voyelles /i, u, a/ extraites de la première syllabe des formes suivantes :

/t+aka+t/ "foyer", /t+afuk+t/ "soleil"
/t+isi+t/ "endroit" surélevé), /t+ifaw+t/ "lumière"
/t+udi+t/ "beurre", /t+usi/ "elle a pris".

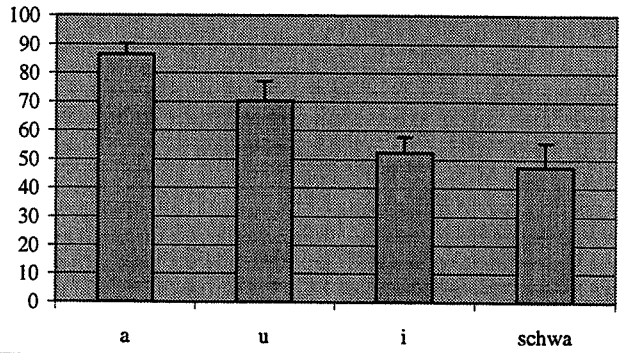


Figure 5: Durées moyennes des voyelles (10 mesures par voyelle) pour le locuteur de Tiznit.

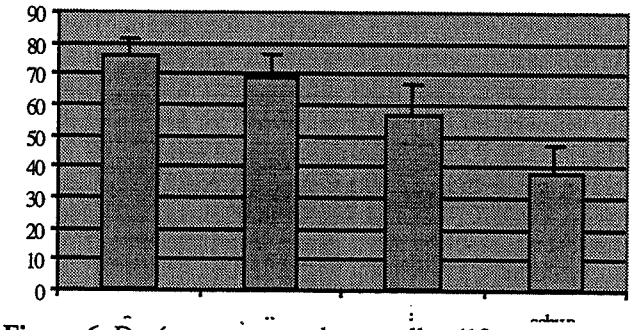


Figure 6: Durées moyennes des voyelles (10 mesures par voyelle) pour le locuteur de Tanalt.

Les figures 5 et 6 reportent pour deux locuteurs la durée moyenne pour ces voyelles périphériques et pour le schwa non transitionnel (valeur supérieure à 20 ms) des tableaux 1 et 2. Il apparaît que le schwa non transitionnel est bien la voyelle la plus brève et le /a/, comme dans la majorité des langues du monde, la voyelle la plus longue. L'écart de moyenne entre le /i/ et le schwa est pour le locuteur de Tanalt de 19 ms et pour le locteur de Tiznit de 5 ms, ce qui constitue un indice de discrimination insuffisant à soi seul. Il est donc vraisemblable que le timbre joue également un rôle corollairement.

4. LE TIMBRE DU SCHWA

Nous avons étudié la dispersion des voyelles dans un espace F1/F2 pour les trois locuteurs qui produisent le schwa. Pour les voyelles /i, u, a/ le corpus précédent était complété par les formes suivantes :

/t+adun+t/	"graisse"
/t+izi+t/	"petite mouche"
/t+uga/	"herbe".

Pour le schwa, nous nous sommes appuyés sur les items cités précédemment. L'ensemble du corpus est constitué de formes non emphatiques. Les enregistrements ont été effectués sur DAT en chambre insonorisée et nous avons réalisé les analyses acoustiques avec le logiciel MultiSpeech. Pour chaque voyelle dix échantillons ont été analysés.

La figure 7 permet de constater qu'il n'y a pas de chevauchement entre les ellipses caractérisant la distribution des voyelles périphériques et du schwa pour chaque locuteur pris séparément. La distribution du /a/ et du /ə/ occupe toutefois la partie centrale de l'espace et se chevauche lorsqu'on superpose comme dans la figure 7 la production de plusieurs locuteurs. Le contraste de timbre pour un locuteur donné constitue un indice complémentaire et sans doute nécessaire à la différence de durée pour établir l'identité perceptuelle du vocoïde.

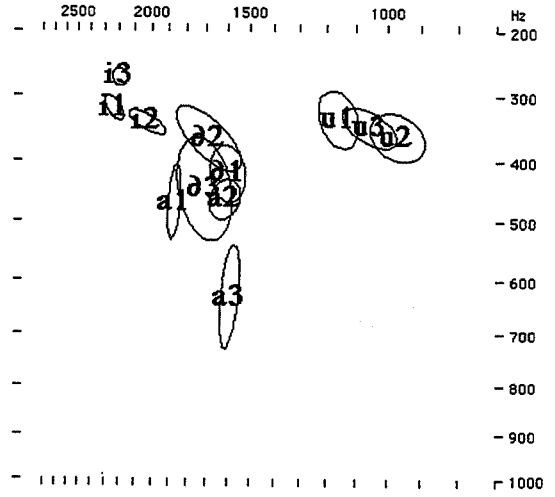


Figure 7: Distribution dans l'espace acoustique F1/F2 du schwa et des voyelles périphériques pour 3 locuteurs

Le parler tachelhit d'Imdlawn, sur lequel reposent les analyses de Dell et Elmedlaoui a été étudié par [Cole 99]. L'auteur conclut que le schwa "épenthétique" est influencé par les caractéristiques des segments adjacents, ce qui est attendu. D'une façon plus surprenante, il pointe sur la possibilité de différencier pour le timbre deux schwas distincts phonétiquement. Les données présentées ici ne permettent pas de confirmer cette interprétation.

5. PERCEPTION DU SCHWA COMME NOYAU

Dans [Puech 99] une étude perceptuelle a été effectuée sur des formes de tachelhit.
Les sujets avaient notamment comme tâche de classer /t+xzn+t/ " tu as emmagasiné " avec des formes comme /ks/ " fais-le paître !" ou /gis/ "dedans" (interprétés comme monsyllabique) ou /t+asa/ "foie" (interprété comme bisyllabique).

Le stimulus comportait un schwa "épenthétique" de 32 ms entre l'occlusive initiale et la fricative uvulaire comme le montre la figure 8.

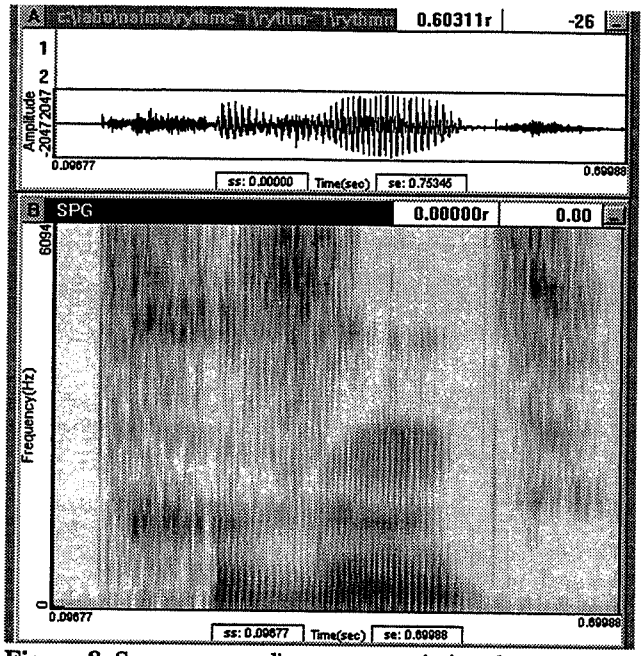


Figure 8. Sonagramme d'une prononciation de *txznt*

Nous avons eu trois types de réponses. Un sujet a classé /txznt/ avec le groupe des formes "monosyllabiques", un autre avec le groupe "bisyllabique". Les trois autres sujets ont varié au cours du même test entre les deux choix.

6. INTERPRÉTATION PHONOLOGIQUE

Dans le cadre des hypothèses faites par [Dell 96], de [Cole 96] et des données présentées dans cette contribution, le vocoïde schwa est susceptible d'être interprété comme :

a) une voyelle de même nature que les voyelles périphériques,
b) une interprétation phonétique du voisement phonologique d'une consonne,
c) l'implémentation par une voyelle centralisée à durée variable d'un noyau syllabique.

La première hypothèse peut être rapidement écartée : les voyelles périphériques /i, u, a/, qu'elles jouent, en berbère, un rôle radical au même titre que les consonnes ou qu'elles aient un statut affixal ne sont pas élidables et ne sont donc pas susceptibles d'une alternance avec une réalisation zéro. La deuxième renvoie au modèle de Dell et Elmadlaoui, qui relient phonologiquement un vocoïde transitionnel au voisement d'une consonne. Ainsi dans une forme comme [təgni], le schwa se rattache pour eux au voisement de l'occlusive.

Ce modèle prédit l'impossibilité d'un schwa entre deux consonnes non voisées. Cette prédiction est compatible avec les données présentées pour le locuteur de Massa mais pas avec celles des trois autres locuteurs, ni avec le cas de /t+xzn+t/ commenté en section 4.

La dernière hypothèse a été, sous des formes diverses, traitée par Coleman. Dans sa contribution de 1996, il montre qu'un schwa "épenthétique" réalise un noyau syllabique non occupé par une voyelle lexicale.

Cette approche permet une lecture phonologique des figures 1 à 3. La figure 1 réalise une séquence CVC avec une correspondance terme à terme entre une séquence phonétique et phonologique.

La figure 2 compresse le vocoïde, qui contrairement à une voyelle lexicale n'a pas à exprimer phonétiquement une "couleur" symbolisable par une particule {I}, {A}ou {U}. La figure 3 illustre le cas où une friction plus intense de la consonne continue non voisée se substitue à la réalisation du vocoïde.

CONCLUSION

En berbère seules les formes verbales ne comportent pas obligatoirement de voyelles lexicales, /i, a, u /. Le schwa, voyelle centralisée et la plus brève, est généralement présente dans les formes où les autres morphosegments sont tous consonantiques. Mais le tachelhit permet aussi aux consonnes voisées et aux consonnes continues non voisées de couvrir le centre de syllabe. On a dès lors une variabilité qui est fonction des locuteurs et de la complexité consonantique des formes et du débit. La question reste ouverte pour les occlusives non voisées. Dans le modèle phonologique auquel nous adhérons pour le tachelhit, toute forme comporte un élément V qui, en l'absence d'une particule colorée attachée, se réalise ou non en schwa.

BIBLIOGRAPHIE

[Cole96] Coleman John (1996) "Declarative Syllabification in Berber Tashlhiyt", P. 175-216, In *Current Trends in Phonology : Models and Methods*, (Jacques Durand & Bernard Laks (eds), volume 1, CNRS ESRI, Paris X.

[Cle97] Clements G.N. (1997), Berber Syllabification : Derivations or Cons-traints ?, p. 289-330, *in Derivations and Constraints in Phonology* (Iggy Roca (ed.), Clarendon Press, Oxford.

[Cole99] Coleman John (1999) "The nature of vocoids associated with syllabic consonants in Tashlhiyt Berber, *Proceedings of the 14 International Congress of Phonetic Sciences*, San Francisco, 1-7 Août 1999, P.735-738.

[Dell96] Dell, F & M. Elmedlaoui, (1996) "Nonsyllabic transitional vocoids in Imdlawn Tashlhiyt Berber", P. 217-243, In *Current trends in phonology : models and methods*, (J. Durand & B. Laks (eds), volume 1, CNRS ESRI, Paris X.

[Lou96] Louali N. & G. Puech (1996) "Syllabic consonants in Tashlhit Berber : the case of unvoiced stops", communication au colloque sur : *The Phonology of the World's Languages: The Syllable*, Pezenas, 21 au 24 Juin.

[Puech99] Puech G. & N. Louali, "Syllabification in Berber : the case of Tashlhiyt", *Proceedings of the 14 International Congress of Phonetic Sciences*, San Francisco, 1-7 Août 1999, p. 747-750.

[Zec95] Sonority constraints on syllable struc-ture, *Phonology* 12, p. 85-129.

Sur la glottalisation et l'occlusive glottale en persan

Assadi Sh. S.

Université de Paris III (Sorbonne Nouvelle)
ILPGA (Institut de Linguistique et Phonétique Générales et Appliquées)
Tel :01 43 26 37 80 Fax : 0144 43 05 73

E mail (assadisuzanne27@hotmail.com) ou (assadi@msh-paris.fr)

ABSTRACT

This study analyzes the physiological (laryngography and pneumatography) and acoustic aspects of glottalization and the glottal stop in isolated words and sentences in Persian.

The laryngographic data generally shows a closing of the glottis before the beginning of the word initial vowel. The air flow curve indicates an interruption in the air flow which corresponds to the closure of the glottis. The production of a glottal stop in the middle or at the end of a word may range from a partial constriction of the glottis to a complete closure in the case of isolated and emphatic words. Acoustic study of glottalization is characterized by an irregular vibration of the vocal folds and a lowering in pitch and amplitude ; glottal closure is characterized by silence.

The study of the influence of speaking rates of two subjects showed a 24% and a 11% increase in glottalization for a slow rate as opposed to a fast one.

The speakers have in common the presence of glottalization between any two vowels separated by a morphological boundary.

1. INTRODUCTION

L'occlusive glottale est produite par l'adduction des cordes vocales. Selon que la tenue de la voyelle est prononcée avec un coup de glotte, c'est-à-dire un blocage brutal des cordes vocales ou non, on distingue les glottalisées et les non glottalisées. En comparant les premières périodes au début de la voyelle, on peut constater que le mode d'attaque peut varier selon les langues. Les voyelles allemandes ont une attaque forte, qui peut sembler agressive et les voyelles françaises, une attaque douce.

La glottalisation est un phénomène général dans toutes les langues où son utilisation semble marquer la voyelle initiale du début de mot, elle peut distinguer un dialecte et signaler le tour de la parole aux autres. Klatt [Kla90]. Les phénomènes de la glottalisation ne sont pas définis clairement dans la littérature. La glottalisation est un terme qui couvre «creak», voix craquée, les phénomènes de laryngalisation, les articulations pulmoniques accompagnées de la fermeture glottale ainsi que les éjectives et les implosives.

Ladefoged [Lad73] propose les différentes configurations glottales sous le terme de « glottal stricture », rétrécissement de la glotte, allant du plus fermé au plus ouvert:

1. "glottal stop" occlusive glottale ou coup de glotte
2. "creaky voice", voix laryngalisée
3. « stiff voice » faible degré de laryngalisation
4. « voice »; voisé, sonore
5. « slack voice", faible degré de « breathy voice" voix soufflée
6. « breathy voice », voix soufflée ou voix murmurée
7. « voiceless », non voisé, sourd
8. « spread », la production de sons aspirés.

« Creak » est une sorte de vibration irrégulière dans laquelle chaque pulsation glottale devient audible.

Dans la voix laryngalisée « creaky voice » qu'on appelle aussi voix craquée et «vocal fry», les aryténoïdes sont très rapprochés et en même temps avancés; une partie des cordes vocales est fermée, tandis que l'autre partie peut vibrer. A cause de l'avancement des aryténoïdes, les cordes vocales ont tendance à être moins tendues et sont donc susceptibles de vibrer à des fréquences plus basses. En "creaky voice", il y a plus d'adduction des cordes vocales que dans la voix modale, et il existe une combinaison du voisement et du « creak ».

D'après Catford [cat77], la fermeture de la glotte pour la production du "creak" est moins importante que celle de l'occlusion complète, pour l'occlusive glottale. Le mécanisme précis du "creak" n'est pas complètement évident. Les cordes vocales sont en contact mais pas très tendues et l'air échappe à travers une petite aperture proche de la partie avant des cordes vocales, ce qui produit un son claquant. La pression sous glottique est très basse. Le "creak" a une fonction phonatoire, tandis qu'une contraction plus importante des cordes vocales crée l'occlusive glottale, qui n'a pas une fonction phonatoire, mais articulatoire.

Laver (1980) définit la forme neutre de la phonation dans laquelle la vibration de vraies cordes vocales est périodique et sans friction audible, comme la référence et à partir de celle-ci décrit les autres types de phonation. Selon lui, les deux termes (laryngalisation et glottalisation) concernent la constriction glottale (glottal stricture).

La laryngalisation joue un rôle phonologique dans beaucoup de langues. Dans les langues à ton, les syllabes à tons bas ou à ton descendant sont phonétiquement caractérisées par un creak ou voix craquée.

Selon Pierrehumbert [Pie92], en anglais, il y a une occlusive glottale au début de la voyelle initiale du mot, surtout quand elle est en initiale, après une pause ou bien quand elle suit un mot se terminant par une voyelle.

Fischer Jorgensen [Fis89] fait une analyse acoustique et physiologique du stod danois. D'après elle, du point de vue phonologique, le stod est une marque prosodique pour définir la syllabe dans certains types de mot. Du point de vue phonétique, c'est une sorte de phonation reliée à la voix craquée.

Kohler [Koh94] étudie la glottalisation et l'occlusive glottale en allemand. En ce qui concerne la catégorie du mot initial, après le silence, la présence d'une occlusive glottale est plus générale que son élision. Selon ses termes, une attaque douce avec ou sans glottalisation est rare.

1 LA GLOTTALISATION ET L'OCCLUSIVE GLOTTALE EN PERSAN

Selon la littérature, phonétiquement, les voyelles initiales du mot en persan sont précédées d'un coup de glotte.

-Du point de vue phonologique, selon que les chercheurs attribuent ou non un statut de phonème à l'initiale au coup de glotte, la structure syllabique peut interpréter différemment.

Certains chercheurs proposent sept types de syllabe en persan (v- vc- vcc- cv- cvc- cvcc- cvccc). Scott [Sco57] en propose quatre, la syllabe commençant toujours avec une consonne.

-Historiquement, l'occlusive glottale du persan correspond à la pharyngale sonore et au coup de glotte arabe. Selon Sadeghi [Sad75], en finale et à l'intérieur du mot, il est distinctif et ne se trouve que dans les mots d'origine arabe. Mais d'après lui, rien ne nous incite à attribuer à une influence arabe son apparition en position initiale avant la voyelle. Le persan a pu facilement accepter ce phonème pour deux raisons: premièrement, parce que la langue possédait déjà un phonème /h/ qui aurait fonctionné comme l'amorce d'un ordre glottal. Deuxièmement, comme le coup de glotte existait, phonétiquement, à l'initiale devant la voyelle, sa prononciation ne posait guère de difficulté. Selon Piowics [Pis85], à l'initiale, il est prévisible mais n'est pas un trait distinctif, mais délimitatif, signalant la présence d'une voyelle précédée d'un silence. Les différentes descriptions du coup de glotte viennent de la variation stylistique.

Les études antérieures sur la glottalisation en persan sont principalement des études perceptives, à l'exception de celle de Gaprivdashvili [Gap64] qui a effectué une étude physiologique et acoustique sur tous les phonèmes du persan et a étudié le coup de glotte au milieu du mot isolé.

2 ETUDE PHYSIOLOGIQUE ET ACOUSTIQUE DU COUP DE GLOTTE ET DE LA GLOTTALISATION EN PERSAN

Nous avons étudié la glottalisation en position initiale, médiane et finale. Le corpus de l'étude physiologique contient des phrases et des mots isolés. En ce qui concerne la position initiale, nous nous sommes inspirée des études d'Umeda [Ume 78] concernant l'anglais et celle de Kohler [Koh 94] en allemand en ajoutant une étude physiologique. Ces études ainsi que celle de Pierrehumbert [Pie92] insistent sur l'irrégularité des vibrations des cordes vocales et l'aspect perceptive de la glottalisation.

2.1 Etude physiologique

L'analyse physiologique comprend une étude pneumotachographique destinée à étudier le débit d'air et une étude laryngographique, pour vérifier le mode phonatoire des cordes vocales.

L'étude a été effectuée au laboratoire de phonétique de l'Université de Paris III. Nous avons essayé d'effectuer une étude laryngographique chez quatre personnes mais elle n'a fonctionné que pour une seule. Un mode d'accolement pas très dynamique des cordes vocales (Gautheron, discussion personnelle), et de la graisse sur le cou (Schoentgen discussion personnelle) sont probablement les raisons de l'échec de trois locuteurs sur quatre.

Le laryngographe nous renseigne indirectement sur les états des cordes vocales sans perturber la phonation. Il y a une bonne correspondance entre les traits du signal EGG et les événements vibratoires des cordes vocales montrés par les radio films. Fourcin [Fou74]. Les impulsions acoustiques générées par les cordes vocales sont prélevées par les électrodes directement à la surface de l'espace glottique. Les variations d'impédance se traduisent par une descente de la ligne de base quand la glotte est ouverte, et par une remontée quand la glotte est fermée.

L'étude laryngographique indique, en général, une fermeture glottale avant le début de la voyelle initiale du mot aussi bien dans les phrases que dans les mots isolés. La courbe du débit d'air montre l'interruption d'air qui correspond à la fermeture glottale (Figure 1). La voyelle commence avec l'ouverture de la glotte, et la descente de la ligne de base sur le tracé traduit par la présence de l'air égressif, (sortant de la bouche).

On voit ici que les variations d'un phénomène physiologique sont intégrées dans un système phonétique. La réalisation du coup de glotte au milieu du mot (les mots arabes) a plusieurs degrés: on peut avoir un rétrécissement de la glotte jusqu'à une fermeture glottale complète.

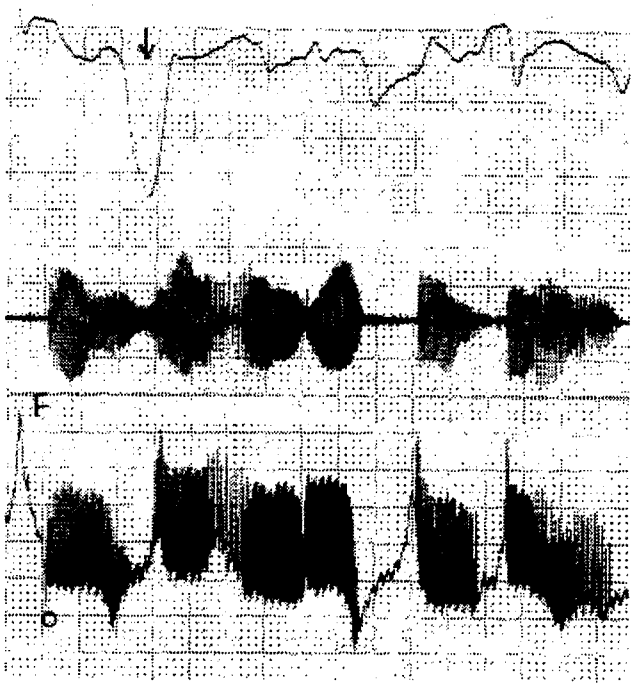


Figure 1 : Etude physiologique de la voyelle initiale du mot dans la phrase « *anha anra* xordeand » (en haut le pneumotachogramme, au milieu le signal et en bas le laryngogramme). F= Fermeture de la glotte. O= Ouverture de la glotte.

2.2 *Etude acoustique*

2.2.1 L'occlusive glottale et la glottalisation selon le débit (lent et rapide)

2.2.1.1 En position initiale du mot

216 voyelles initiales au début du mot ont été étudiées dans un texte lu, en débit lent et rapide. Les voyelles ont été classées selon les trois catégories ci-dessous :

Type A : L'occlusive glottale suivie de la glottalisation.

Type B : La glottalisation (de deux voyelles successives séparées par une frontière morphologique (v-v) (ou sonnantes- voyelle)). L'irrégularité de l'onde, l'abaissement de l'amplitude et les striations verticales sur le spectrogramme.

Type C : L'irrégularité des cycles.

Le choix du classement a été fait par l'examen des premiers cycles au début de la voyelle (figure 2), ainsi que par la courbe de mélodie et d'intensité et du point de vue perceptive.

Le début de la voyelle est marqué en général, par quelques cycles irréguliers (2 à 7 cycles), suivis des vibrations périodiques régulières. Le tableau 1 montre le pourcentage

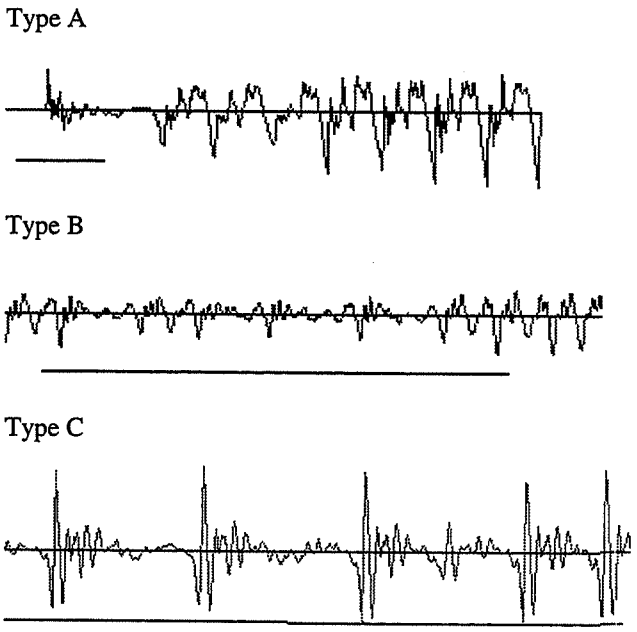


Figure 2 : les différentes formes du signal au début de la voyelle initiale du mot.

de l'occurrence par chaque cas. En débit rapide, l'occurrence du coup de glotte (type A) est respectivement de 24% et de 11% de moins pour le sujet A et B qu'en débit lent. De plus, l'enchaînement des mots (%11 pour le sujet A et %19 pour le sujet B) aboutit à la disparition du coup de glotte. Ce qui suggère un changement de coupe syllabique, par exemple « ?az / ?an » -> « ?a / zan. ».

Tableau 1: Le pourcentage de l'occurrence de chaque type de glottalisation. « Autres » indiquent les cas douteux, semi- voyelles et l'enchaînement des mots).

	Loc. A		Loc. B	
	lent	rapide	lent	rapide
Type A	43%	19%	26%	15%
Type B	50%	39%	31%	28%
Type C	6%	17%	33%	26%
Autres	2%	26%	9%	31%

Chez les deux sujets, l'extension de la glottalisation varie en fonction de l'accent d'insistance.

Ce qui est en commun chez les deux sujets est la présence de la glottalisation entre deux voyelles successives séparées par une frontière morphologique. La glottalisation (v-v ou sonnantes- v) est marquée par l'irrégularité de l'onde et les striations verticales sur le spectrogramme, ainsi que par l'abaissement de l'intensité et de la F0.

2.2.1.2 *En position médiane et finale du mot*

Dans cette partie, 72 mots avec le coup de glotte en position médiane (tous précédés d'une voyelle) ainsi que 16 mots avec le coup de glotte en position finale ont été

étudiés dans les deux débits. Selon le style soigné ou rapide, on peut avoir une fermeture glottale complète ou la glottalisation de la voyelle précédente. En débit lent avec un style très soigné, pour le sujet B 39% des cas ont une occlusion complète contre 6% en débit rapide. Ce pourcentage est respectivement de 44 et de 17 pour le sujet A. La durée de l'occlusion est parfois jusqu'à 700 ms. En débit rapide, l'occlusion complète est moins fréquente et dans la majorité des cas, la glottalisation est marquée par des cycles irréguliers et un abaissement de l'amplitude.

En ce qui concerne la position finale du mot, dans les deux débits aucun sujet n'a réalisé l'occlusive glottale. Elle est réalisée seulement en mots isolés prononcés avec un style très soigné.

2.3 *Etude de phrases*

Les mots isolés ont été insérés dans des phrases (10 phrases ont été répétés trois fois par deux sujets). En premier lieu, pour vérifier si la glottalisation dépend d'une voyelle précise, nous avons choisi les mots terminant par les voyelles /i e *a* o u/ alors que le mot suivant commençait toujours par la voyelle */a/*. Nous avons constaté que dans tous les cas, la voyelle suivante est glottalisée. La variation intra- individuelle est au niveau de l'extension de la glottalisation.

Afin d'étudier la glottalisation à la frontière du mot ou du morphème, dans le même corpus nous avons étudié des paires de mots qui se distinguaient par une pause ou un coup de glotte, à la frontière (du mot ou du morphème). Exemple /?az/ /ᵖ?*an*/ et /?a*z*a*n*/. Nous avons trouvé que la glottalisation existe uniquement avant la voyelle initiale du mot après le silence.

3 CONCLUSION

L'étude laryngographique montre une fermeture glottale avant le début de la voyelle initiale du mot. La courbe du débit d'air indique l'interruption d'air qui correspond à la fermeture glottale. Dans toutes les positions (début, milieu et fin du mot), plus le style est soigné plus le contact des cordes vocales est long. En débit lent, au milieu du mot, il y a plus de fermeture glottale complète qu'en débit rapide. A la fin du mot l'occlusive glottale n'est réalisée dans aucun débit. Ce qui est en commun chez les deux sujets est la présence de la glottalisation entre deux voyelles successives séparées par une frontière morphologique (v-v). Ces résultats complètent nos connaissances sur la variabilité de la glottalisation dans les langues, Dans le style soigné où l'intelligibilité augmente, l'occurrence du coup de glotte est renforcée. Mais dans la parole synthétisée, le coup de glotte est-il nécessaire pour percevoir l'aspect naturel de la parole ?

BIBLIOGRAPHIE

[Cat77] Catford, J. C. (1977), Fondamental Problems in Phonetics. Edinburgh: Edinburgh University Press.

[Dil95] Dilley, L.C. & Shattuck-Hufnagel, S. (1995), «Variability in glottalization of word onset vowels in American English» Proceedings of the XIIIth International Congress of Phonetics Sciences. Stockholm 4, pp. 586 - 589.

[Fis89] Fischer-Jorgensen, E. (1989), « Phonetic analysis of the stod in Standard Danish ». Phonetica, 46. pp. 1 -59

[Fou4] Fourin, A.M. (1974). « Laryngeal examination of vocal fold vibration » . In B. Wyke (Ed.), *Ventilatory and phonotary control systems* (pp.315-333). London: Oxford University Press.

[Gap64] Gaprivdashvili, Sh. & Giunashvili, J. (1964) The Phonetics of Persian Language. Oriental Institute of the Academy of Science, Georgian SSR.

[Kla90] Klatt, D.H.& Klatt L.C. (1990), «Analysis, synthesis, and perception of voice quality variation among female and male talkers». JASA 87 (2) pp 820-844.

[Koh94] Kohler, K.J. (1994), «Glottal stops and glottalization in German. Data and theory of connected speech processes». Phonetica 51, pp. 38 - 51

[Lav80] Laver, J. (1980), The Phonetics Description of Voice Quality. Cambridge: Cambridge University Press.

[Lad73] Ladefoged. P. (1973), « The features of the larynx. » Journal of Phonetics.1, pp.73- 83.

[Lad88] Ladefoged. P(1988), « A Note on Some Terms for Phonation Types », in: Vocal Physiology. Voice Production, Mechanisms and Functions (Fujimura. O. éd.), Raven press New York pp. 373-375.

[Pie94] Pierrehumbert, J (1994), "Prosodic Effects on Glottal Allophones" in Fujimura, éd. Vocal Fold Physiology 8. Singular Publishing Group. San diego.

[Pis85] Pisowicz A. (1985), Origine of the new and middle persian. Phonological systemes.

[Sad75] Sadeghi, A.(1975), « L'influence de l'arabe sur le système phonologique du Persan ». La linguistique. Volume 11 fasc. 2.

[Sco64] Scott, Ch .T. (1964), « Syllable Structure in Teheran Persian. Anthropological Linguistics ». Volume 6, no. 1. pp. 27 - 30.

[Ume78] Umeda, N. (1978) : « Occurrence of Glottal Stops ». Jasa. vol. 64 n. 1.

Autour de l'harmonie vocalique en français

Philippe Boula de Mareüil[1], *Zsuzsanna Fagyal*[2]

[1] Elan TTS — 4 rue Jean Rodier — F-31400 Toulouse
Tél. : ++33/0 5 61 36 07 77 — Fax : ++33/0 5 61 36 89 11
Mél : mareuil@elan.fr — http://www.elan.fr

[2] *French Department of the University of Illinois at Urban-Champaign*
2090 Foreign Languages Building — 707 S. Mathews Ave. — Urbana, IL 61801
Tél. : ++1 +217 265-0743 — Fax : ++1 +217 244-2223
Mél : zsfagyal@uiuc.edu — http://www.uiuc.edu

ABSTRACT

In this contribution, the realisation of mid vowels in non-word-final syllables is addressed, and their behaviour with regards to vowel harmony is investigated. Focussing on transcription guidelines, the applied methodology and data of two experiments are described: one based on speech recognition and automatic alignment, the other on acoustic analysis. Mid vowels are pronounced according to their underlying representation in an average of (onlu) 60% of all cases.

1. INTRODUCTION

Pour les voyelles intermédiaires /e/, /ɛ/, /o/, /ɔ/, /ø/, /œ/ en français métropolitain, une grande variabilité idio-syncratique et régionale entre en ligne de compte [Mar77], que les règles comme la loi de la position, dérivées uniquement du contexte consonantique, ne permettent pas de modéliser. Dans une publicité célèbre, par exemple, *lait* était prononcé avec un [e] très fermé, témoignant de la tendance actuelle à la neutralisation de l'opposition [e]~[ɛ] en syllabe ouverte [Lef88]. Mais un autre processus phonologique important détermine la distribution de ces voyelles dans le mot : l'harmonie [Oha94]. La règle de l'harmonie vocalique stipule que les voyelles d'un mot ou d'un groupe de mots sont modifiées sous l'influence d'une voyelle apparaissant dans le même domaine, de telle façon que toutes les voyelles présentes dans ce domaine en viennent à partager certains traits. Cette définition générale est restreinte, en français, à l'influence qu'exerce la voyelle de la syllabe accentuée sur l'aperture des voyelles qui la précèdent (en cas d'accent final) ou qui la suivent (en cas d'accent initial) [Tra94]. Des erreurs d'orthographe et autres jeux de mots dans les courriers électroniques témoignent des phénomènes d'harmonie et de neutralisation [Bar99] : **préférable, viendrai* (pour *viendrez*), *NRV* (pour *énervé*), *Monsieur et Madame Pourferlavésèle ont un fils...*, *ces* ou *ses* pour *c'est* (et inversement), **espère, pourrez* (pour *pourrais*), *allez* (pour *allais*), *donner* (pour *donnaient*).

Pour étudier la réalisation des voyelles intermédiaires en français, dans une première expérience (section 2), un système de reconnaissance de la parole a été utilisé [Add99], dont une série d'évaluations a prouvé la fiabilité. Sur un corpus de 35 heures de parole spontanée transcrit orthographiquement (MASK [Lam95], collection de dialogues finalisés, d'information voyageur), des transcriptions phonétiques ont été obtenues en alignant automatiquement les données acoustiques avec un graphe de prononciation, dérivé d'un dictionnaire incluant des variantes phonétiques.

Dans une deuxième expérience (section 3), nous avons comparé les réalisations des voyelles moyennes de 25 mots communs à MASK et à un corpus d'entretiens sociolinguistiques transcrits et analysés à la main. Des mesures acoustiques ont été prises, confirmant dans l'ensemble les premiers résultats : on peut parler non seulement de neutralisation, mais aussi d'harmonie vocalique en français.

2. PREMIERE EXPERIENCE : RECONNAISSANCE DE PAROLE ET ALIGNEMENT AUTOMATIQUE

2.1. Corpus, alignement et représentation des variantes

Les données du corpus MASK constituent 38 000 énoncés provenant de 409 locuteurs sans accent marqué : le vocabulaire est de 2 000 mots (différents). Afin de déterminer la séquence d'allophones réalisée dans un énoncé donné, une chaîne de Markov est formée en concaténant les prononciations associées aux mots dans la transcription orthographique correspondante — les modèles acoustiques sont des modèles de Markov cachés à densité continue. L'espace de recherche est ainsi contraint pour le système. Si des variantes de prononciation sont ajoutées dans le lexique, par des lois phonologiques (comme c'est le cas ici, générées par un convertisseur graphème-phonème), un graphe de prononciation est construit et aligné avec le signal. Dans ce cas, le décodeur produira la prononciation la plus probable, décision binaire pour ce qui nous concerne dans cette étude, entre un timbre fermé et un timbre ouvert.

Un système de transcription très proche de l'API a été utilisé, avec un ensemble de diacritiques et de méta-symboles pour représenter les variantes de prononciation. Dans le cadre d'un dictionnaire électronique de références phonétiques avec variantes [Bou00], des diacritiques ont été introduits, en particulier: /</ et />/ d'une part, et /-/

d'autre part, spécifiant le phonème qui précède, indiquent respectivement l'ouverture et la fermeture d'une part, et les segments « flottants » de l'autre. Ainsi, le /ɔ>/ est un /ɔ/ ouvert qui se ferme, et le /o</ est un /o/ fermé qui s'ouvre.

Afin de trancher par exemple entre /ɔ>/ et /o</, à partir de la transcription sous-jacente, deux règles (et seulement deux) sont précisées pour produire les assimilations de timbre des paires /olɔ/, /elɛ/ et /ølœ/. Elles s'appliquent séquentiellement :

1. Une voyelle ouverte se ferme en syllabe ouverte, hormis en syllabe finale « ferme » (i.e. quand un schwa final est optionnel, comme dans *alcalose*, la voyelle précédente garde son timbre). D'où les transcriptions de *têtu* (/tɛ>ty/) face à *tête* (/tɛtə-/).
2. Il y a assimilation régressive des apertures (i.e. modification au contact du phonème qui suit), par harmonie vocalique : dans une séquence semi-fermée semi-ouverte (resp. semi-ouverte semi-fermée), la première s'ouvre (resp. se ferme). Dans les séquences /ɛ...e, ɛ...o, ɔ...e, ɔ...o/, quelle que soit la complexité du groupe consonantique /.../, les semi-voyelles (ou semi-consonnes) sont transparentes : elles ne gênent pas la propagation.

Cela conduit à :

fêté : /fɛ>te/ (par 1)

testé : /tɛ>ste/ (par2)

microphone : /mikʁo<fɔnə-/
(par 1, qui ouvre le /o/ fermé du préfixe)

jeunesse : /ʒœnɛsə-/ (par 1, qui donnerait /ʒœ>nɛsə-/, puis 2, qui ouvre le /œ>/).

Les phénomènes en cascade tels que la resyllabation résultant de la chute du schwa sont délicats. Dans le cas bien connu de *médecin*, nous avons par exemple /me<də-sɛ̃/.

Nous entendons par « forme sous-jacente » le timbre défini dans la représentation lexicale du mot, et utilisé ici comme point de départ de la dérivation des règles d'harmonie et de neutralisation. La forme sous-jacente du 'o' graphique, ailleurs qu'en syllabe finale « ferme », est /ɔ/, hormis avant consonne allongeante (/z/), et hormis dans les préfixes *psycho-*, *auto-*, etc., où le phonème cible (éventuellement accompagné du diacritique d'ouverture) est /o/ [Wal76]. Pour le 'e' devant consonne double, la règle par défaut est la suivante : /e/ si le 'e' est initial et si la consonne n'est pas 'r' (ex. *ecchymose* /ekimozə-/, *effort* /e<fɔʁ/) ; /ɛ/ dans les autres cas, le plus souvent — même si les choses sont plus compliquées.

2.2. Résultats

Le résultat de l'alignement est une suite de lignes comme:
réservation (1577) : ʁɛzɛʁvasjɔ̃ (933),
ʁezɛʁvasjɔ̃ (644)
Bordeaux (1001) :bɔʁdo (781), boʁdo (320)

ce qui signifie — pour la première — que sur les 1 577 occurrences reconnues du mot *réservation*, le premier 'é' est 933 fois contre 644 plus proche d'un /ɛ/ que d'un /e/. En moyenne (pondérée) sur 192 entrées comprenant au moins un /</ ou un />/ (12 678 occurrences), les voyelles intermédiaires avec diacritiques s'alignent à 60 % (seulement) avec leur timbre sous-jacent : le /ɔ>/ avec le [ɔ], le /e</ avec le [e], etc. Notons que si nous avons presque autant de mots avec /olɔ/ qu'avec /elɛ/, nous n'avons quasiment pas d'occurrence de /ølœ/ (*deuxième*). De façon plus détaillée :

- 50 % des /e</ s'alignent avec des [e] ;
- 59 % des /ɛ>/ s'alignent avec des [ɛ] ;
- 87 % des /o</ s'alignent avec des [o] ;
- 64 % des /ɔ>/ s'alignent avec des [ɔ].

La relative faiblesse du premier chiffre est très influencée par l'importante population des mots en *réserv-* (plutôt ouverts), tandis que le chiffre le plus élevé doit beaucoup au mot *Beauvais* (très porté vers le [o]).

Dans le tableau 1, le [o] de *proposer* reflète la tendance à l'harmonie sous l'effet de l'allophone fermé de la syllabe suivante [o:z]. On remarque en revanche que l'harmonisation fermant le [ɛ] est quelque peu empêchée par la présence d'un /ʁ/ subséquent (ex. *libẹrté*). Quant à la série antérieure, les voyelles labiales dans *deuxième* et *déjeuner* sont, dans leur réalisation, conformes à leur timbre sous-jacent, malgré une certaine incertitude dans le second cas.

Malgré l'orthographe, la voyelle /e/ dans les mots *était*, *période* et *réserve* prend le timbre mi-ouvert de la voyelle accentuée qui suit. Ceci contredit [Lef88] qui constate, d'après l'analyse d'un corpus télévisé, que la présence d'une marque explicite de prononciation mi-fermée (ici l'accent aigu) empêche l'ouverture de la voyelle.

3. DEUXIEME EXPERIENCE : ANALYSE ACOUSTIQUE

Par la suite, nous avons fait l'analyse acoustique des voyelles moyennes de 25 mots communs à MASK et à un corpus d'entretiens enregistrés avec 2 hommes et 2 femmes de la région parisienne (âgés de 25 à 35 ans). Nous voulions savoir si les tendances observées dans MASK se confirment ou non par nos mesures.

3.1. Corpus sociolinguistique

Les locuteurs ont été enregistrés à Paris ou aux États-Unis (en courte visite), entre 1997 et 1999 dans le cadre d'une enquête sociolinguistique plus large, conçue selon la technique des entretiens laboviens [Lab78]. Chaque entretien est composé de différents modules, tels que récits, conversations, exercices à trous et lecture de texte, qui visent à saisir, pour chaque locuteur :

- le style formel (mot-cible lu pour la première fois sur la carte tendue au sujet, toute lecture subséquente de mots, lecture de textes) ;

- le style informel (récits, conversations, définition du sens d'un mot, énumérations spontanées).

Les mots analysés dans cette comparaison ne constituant pas le cible de l'enquête, nous n'avons pu en étudier qu'une vingtaine, apparaissant de façon aléatoire dans l'un ou l'autre (plus rarement dans chaque) entretien. Les mesures acoustiques ont été prélevées au milieu de la section stable de la voyelle. L'identité de chaque voyelle a été déterminée, pour chaque locuteur, en fonction des zones de dispersion des voyelles intermédiaires dans des paires minimales.

3.2. *Mots-cibles dans MASK et les entretiens*

Tableau 1 : mots communs à MASK et au corpus d'entretiens.

	[ɛ]	[e]	[œ]	[ø]	[ɔ]	[o]
aimé (11)	5	6				
aura (15)	-	-	-	-	14	1
chocolat(10)	-	-	-	-	10	
chocolat	-	-	-	-	10	
comment (10)	-	-	-	-	2	8
dernier(194)	165	29				
donne(r)(z)(é) (108)	-	-	-	-	77	38
déjeuner(15)	1	14	8	7		
deuxième (1629)	1629	-	268	1361		
était (2058)	1785	273				
était	2058					
faudrait(26)	26	-	-	-	6	20
fermées (14)	13	1				
horrible(17)	-	-	-	-	12	5
isolés (18)	-	-	-	-	15	3
liberté(186)	186					
payer(z) (161)	96	65				
période (14)	9	5	-	-	14	
possibilités (15)	-	-	-	-	9	6
possible(s) (785)	-	-	-	-	445	340
priori (44)	-	-	-	-	38	6
prochain(s) (109)	-	-	-	-	24	85
proposer(z) (24)	-	-	-	-	-	24
proposer(z)	-	-	-	-	-	24
réserve (1813)	1106	707				
réserve	1813	-				
restaurant (630)	630	-	-	-	560	70
social (114)	-	-	-	-	71	43

Le tableau 1 indique le nombre total d'allophones ouverts et fermés (32 voyelles soulignées) dans 25 mots relevés du corpus MASK. Conformément aux prédictions de 2.2, environ 60 % des voyelles (19 / 32) sont conformes à leur timbre sous-jacent.

Dans le tableau 2 figurent les 11 mots sur 25 dont la prononciation diffère de celle observée dans MASK. Le nombre d'occurrences par locuteur étant très variable, les « + » ne font qu'indiquer des tendances générales.

La voyelle orthographique 'é' dans *était*, *période* et *réserve*, ainsi que le digramme 'ai' dans *aimé* se prononcent exclusivement avec un timbre mi-fermé par les deux locuteurs qui ont prononcé ces mots. Dans les trois premiers mots, la convention orthographique semble déterminante, alors que la prononciation [eme] dans *aimé* serait due à l'harmonie.

Tableau 2 : mots dont la prononciation diffère entre les deux corpus.

	[ɛ]	[e]	[œ]	[ø]	[ɔ]	[o]
aim(é)(r)(z) (6)		+				
comment (20)			+		+	
donner(z)(é) (14)			+		+	
était (100)		+				
était (100)	+	+				
paye(é)r(z) (7)		+				
période (6)		+				
possibilité(s) (6)					+	
possible(s) (8)					+	
réserve(é).. (4)		+				
social (6)			+		+	

Les voyelles antérieures dans *payer* et *était*, résultats d'une harmonie de directions opposées, sont perçues plutôt comme des [e]. Mais du point de vue acoustique, elles occupent un espace intermédiaire entre [ɛ] et [e] chez 3 locuteurs sur 4 — pour chaque locuteur, une dizaine d'occurrences ont été affichées (figure 1). Ceci est également le cas lors de la neutralisation des deux voyelles en syllabe ouverte [Fag00].

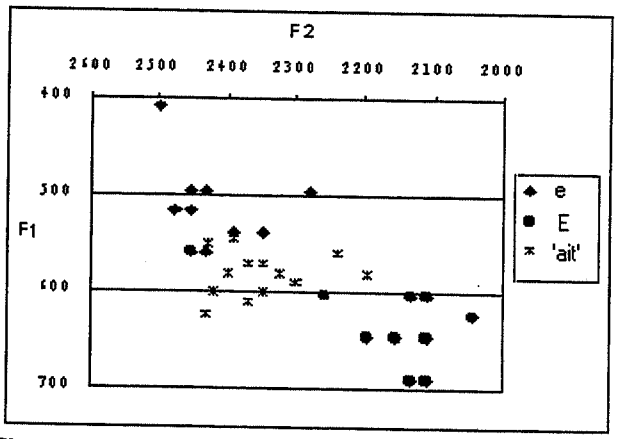


Figure 1 : valeur des formants, dans un plan F1-F2, de 15 occurrences du mot *était* (style informel) superposés à l'espace vocalique maximal de [ɛ] et [e] de Camille M. (33 ans), Paris 18e. Les mesures de [ɛ] et [e] sont des paires minimales *prêt~pré*, *taie~thé*, etc.

On note non pas un changement d'aperture, mais une antériorisation de /ɔ/ vers [œ] dans les disyllabes *comment*, *donne(r)(z)(é)* et *social(e)* chez certains locuteurs (figure 2). Autrement dit, l'harmonie par aperture est supplantée par un phénomène d'un autre ordre : la neutralisation de la distinction entre les traits antérieur~postérieur dans la série labiale mi-ouverte. Ce phénomène avait déjà été relevé par

[Mar69] et étudié plus récemment par [Mal95]. À notre connaissance, le rôle du contexte consonantique dans les cas d'harmonie vocalique n'a pas encore été étudié en français. Quant aux cas de neutralisation, nous renvoyons à l'étude [Mal95].

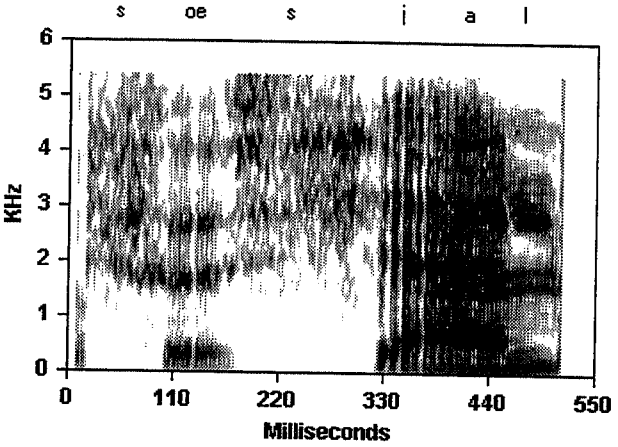


Figure 2 : spectrogramme du mot *sociales* chez Mathilde P. (26 ans), Achères-Ville,Yvelines. Les valeurs élevées et la quasi-équidistance des formants en dessous de 3 kHz de la voyelle [ɔ] (F1=0,54 kHz ; F2=1,7 kHz ; F3=2,7 kHz) témoignent de l'antériorisation de la voyelle en [œ].

4. Conclusion

Dans une première expérience, utilisant la reconnaissance vocale et l'alignement automatique, nous avons observé la réalisation des voyelles intermédiaires non accentuées en français. Il se révèle que celles-ci ne conservent qu'à 60 % en moyenne leur timbre sous-jacent. Les résultats dépendent bien sûr des prononciations autorisées, des modèles acoustiques et plus généralement des paramètres du système de reconnaissance : ils doivent donc être nuancés. Mais globalement, les prédictions fondées sur le corpus MASK ont été confirmées par un autre corpus (une enquête sociolinguistique dans laquelle le contrôle de l'âge et de la provenance des locuteurs est plus étroit) et par son analyse acoustique.

Par ailleurs, le système de transcription ici adopté, quoique plus précis que l'API strict, ne permet pas de distinguer facilement si une voyelle semi-ouverte en syllabe ouverte se ferme par harmonie vocalique ou non. Il ne permet pas non plus d'évaluer l'influence de la morphologie dérivationnelle, par exemple pour le passage de *clair* ([klɛʁ]) à *éclairé* ([eklɛʁe]). Néanmoins, un tel examen objectif de la réalisation des voyelles intermédiaires en français (parisien) peut trouver des applications en synthèse et en reconnaissance de la parole, ainsi qu'en phonétique descriptive, didactique ou corrective. Il conviendrait également qu'il soit étendu à des mots outils comme *les*, et à des locuteurs du « Sud ».

5. Remerciements

Une partie des réflexions exposées ici a été menée au sein d'un programme « Ingénierie des langues » du CNRS, avec C. d'Alessandro, F. Yvon, V. Aubergé et J. Vaissière. Chaleureusement, merci à eux tous, ainsi qu'à M. Adda-Decker (LIMSI), qui a fourni les alignements.

Bibliographie

[Add99] Adda-Decker M., Boula de Mareüil P. & Lamel L. (1999), « Pronunciation variants in French: schwa & liaison », *ICPhS*, San Francisco, pp. 2239-2242.

[Bar99] Barone C. (1999), « Un esame delle realizzazioni vocaliche in francese attraverso la "scrittoralità" della comunicazione in rete », *10e Giornate di Studio del Gruppo di Fonetica Sperimentale*, Naples (à paraître).

[Bou00] Boula de Mareüil P., Yvon, F., d'Alessandro C., Aubergé V., Vaissière J. & Amelot A. (2000), « A French phonetic lexicon with variants for speech and language processing », *LREC*, Athènes (à paraître).

[Fag00] Fagyal Z. (2000), « Les voyelles antérieures moyennes en syllabe ouverte à Paris. », *Proceedings of the Association for French Language Studies Meeting*, Québec (à paraître).

[Lab78] Labov W. (1978), *Sociolinguistic patterns*, University of Philadelphia Press, Philadelphia.

[Lam95] Lamel L., Rosset S., Bennacef S., Bonneau-Maynard H., Devillers L., Gauvain J.-L. (1995), « Development of Spoken Language Corpora for Travel Information », *Eurospeech*, Madrid.

[Lef88] Lefèvre A. (1988), « Les voyelles moyennes dans le français de la radio et de la télévision », *La Linguistique*, vol. 24, fasc. 2, pp. 75-77.

[Mal95] Malderez I. (1995), *Contribution à la synchronie dynamique du français contemporain: le cas des voyelles orales arrondies*, Thèse de doctorat de l'université de Paris VII, Paris.

[Mar69] Martinet A. (1969), « C'est jeuli le Mareuc. », *in Le français sans fard*, PUF, Paris, pp. 191-208.

[Mar77] Martinet A. & Walter H. (1977), *Dictionnaire de la prononciation française dans son usage réel*, France-Expansion, Paris.

[Oha94] Ohala J.J. (1994.), « Towards a universal, phonetically-based, theory of vowel harmony », *ICSLP*, Yokohama, pp. 491-494.

[Tra94] Tranel B. (1994), *The Sounds of French: an Introduction*, Cambridge University Press.

[Wal76] Walter H. (1976), *La dynamique des phonèmes dans le lexique français contemporain*, France-Expansion, Paris.

[Wal92] Walter H (1994), « Les fluctuations mettent-elles en danger une opposition phonologique ? », *La Linguistique*, vol. 28, fasc. 1, pp. 59-68.

Marseillais et Toulousains gèrent-ils différemment leurs pieds ? Caractéristiques prosodiques du schwa dans les parlers méridionaux

Annelise Coquillon, Albert Di Cristo & Michel Pitermann

Laboratoire Parole et Langage, CNRS ESA 6057
Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence, CEDEX 1 FRANCE

ABSTRACT

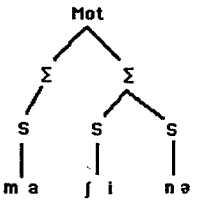
Several studies showed that languages and dialects can be identified by prosodic cues. In a previous work, the first author presented evidence that French native speakers from the South and the North of France could be distinguished by prosodic features [Coq96]. The present paper describes an experiment aimed to characterise two southern French regional accents. For each accent, three speakers read a dialogue. Syllable and vowel nucleus duration was measured inside a prosodic foot containing a schwa at the end of tonal units. The results suggest that the analysed foot was a unit of temporal organisation. This distribution of duration within the foot was sufficient to categorise the two classes of speakers, though it was also influenced by intonation contexts as well as intrinsic vowel duration.

1. INTRODUCTION

L'étude des éléments prosodiques du langage relève à la fois de la recherche des universaux et de celle des propriétés linguistiques des langues individuelles. Il est admis, dans cette double perspective, que les similitudes prosodiques observées entre les langues n'occultent en aucune manière le fait qu'elles possèdent des caractéristiques prosodiques qui leurs sont propres. Plusieurs travaux ont montré qu'un signal de parole réduit à son expression prosodique (énoncés filtrés) contient l'information nécessaire à l'identification de la langue qu'il véhicule (Maidment [Mai83], Le Besnerais [LeB96]). Par ailleurs, on a pu observer récemment (Hirst & Di Cristo [Hir98]) que les différences prosodiques qui affectent les variantes dialectales d'une même langue sont parfois plus accusées que celles qui distinguent deux idiomes différents. Bien que l'on s'accorde sur le rôle joué par la prosodie dans l'identification des dialectes et des parlers régionaux, peu d'investigations expérimentales n'ont été consacrées jusqu'à présent à cette problématique. Néanmoins, Mora [Mor96] a montré que les indices prosodiques pouvaient suffire à distinguer des dialectes espagnols parlés au Venezuela. Des résultats similaires ont été obtenus par Coquillon [Coq96] en ce qui concerne la distinction de locuteurs français méridionaux et non-méridionaux avec des scores supérieurs à 70%. Le travail présenté ici a pour objectif d'affiner cette analyse pour le français méridional en focalisant sur l'une des caractéristiques prosodiques (temporelle en l'occurrence) du schwa final de deux parlers régionaux du Sud de la France : Le français parlé à Marseille et son équivalent toulousain.

2. EXPOSÉ DE LA PROBLÉMATIQUE

Il est bien connu que la prononciation quasi-systématique du schwa représente l'un des traits particuliers de la prononciation du français méridional (Lucci [Luc83], Coquillon [Coq96]). Sa réalisation fortement marquée en fin d'Unité Intonative (U.I.) et d'énoncé le différencie nettement de la voyelle d'appui que l'on rencontre fréquemment dans les mêmes contextes en français parisien. Sa réalisation pleine à la finale d'une U.I. a pour effet de favoriser l'émergence d'une syllabe non-accentogène, rattachée à la syllabe accentuée qui précède pour former une unité supra syllabique équivalente à la notion de pied prosodique ($\sum$), telle qu'elle est conçue par Selkirk [Sel77]. C'est ainsi que le mot « *machine* » aura la représentation arborescente suivante :



Le pied prosodique qui, selon cette conception, peut être formé d'une syllabe unique ou de deux syllabes (notées "S"), est le domaine de gestion des caractéristiques temporelles de ses éléments constitutifs. Un schwa final est ainsi toujours associé - et dépend - de la syllabe précédente avec laquelle il forme un pied prosodique.

Comment cet allongement final est-il régi au sein des pieds prosodiques de structure [cvcə] qui sont attestés dans les deux variétés de français méridional qui nous intéressent ? L'écoute de locuteurs représentatifs de ces deux variétés régionales donne l'impression que les locuteurs toulousains donnent « plus de poids », notamment en ce qui concerne la durée, à la syllabe finale dotée du schwa que les locuteurs marseillais, ce qui semble, par exemple, être particulièrement sensible dans l'élocution du chanteur Claude Nougaro. Cet article présente une vérification expérimentale de l'hypothèse selon laquelle la gestion de l'organisation temporelle au sein des pieds métriques comportant un schwa en position finale d'Unité Intonative, constituerait l'une des

caractéristiques prosodiques qui contribuent à différencier ces deux variétés de français méridional.

3. MÉTHODE

Les deux variétés méridionales ont été sélectionnées en fonction de critères linguistiques. Nous pensons en effet que "l'accent" du Midi provient du substrat occitan (Seguy [Seg50]). Celui-ci comprend plusieurs dialectes : l'occitan moyen, entre autres, est lui-même composé de sous dialectes : le provençal dans la région du Sud-est et le languedocien dans le Sud-ouest, ces deux variétés étant très proches. Ainsi, la région marseillaise est représentée par les anciennes limites du provençal maritime, et la région toulousaine par celles du languedocien méridional.

Le corpus est composé de brefs dialogues simulés, à mi-chemin entre la lecture d'énoncés isolés et la parole spontanée. Ce type de corpus, représentatif d'une interaction "feinte", permet de préserver au maximum les caractéristiques naturelles des accents régionaux, tout en permettant de contrôler certaines variables. Chaque locuteur (qui prononce la totalité des dialogues, interprétant tour à tour le locuteur A puis le locuteur B) réalise une moyenne effective de 37 schwas finaux (dans des mots pour la plupart trisyllabiques situés en fin d'Unité Intonative). Un test préliminaire a permis de sélectionner six locuteurs particulièrement représentatifs de l'une ou l'autre région (trois pour chacune) en fonction des critères suivants : accent toulousain ou marseillais typique et fortement marqué. Un second test contenant uniquement les mots retenus a permis d'établir le fait qu'ils contiennent l'information nécessaire à la discrimination des deux variétés méridionales.
L'enregistrement du corpus a ainsi été réalisé par paires de locuteurs en chambre anéchoïque, puis numérisé à 16 kHz (résolution 16 bits). La segmentation a été effectuée manuellement grâce au logiciel MES (Message Éditeur de Signal, Espesser [Esp96]). Tous les mots contenant un schwa final d'U.I. ont été étiquetés en syllabes et constituants syllabiques (attaque, rime, coda), la segmentation s'effectuant en début et fin de chaque phonème. Les syllabes du pied métrique étaient la syllabe tonique (St) et la syllabe contenant un schwa (Sə). Les mesures ont fait l'objet de validations statistiques par analyse de la variance (Anova) avec un modèle de type linéaire à effet fixe. Nous avons considéré comme seuil critique la valeur de p=0,05.

4. RÉSULTATS

Nous présentons dans cette section l'ensemble des résultats issus de la comparaison des durées des syllabes (4.1) et des noyaux vocaliques (4.2) du pied prosodique final d'U.I., pour les deux groupes de locuteurs.

4.1 Répartition des durées syllabiques

La mesure de la répartition des durées entre les deux syllabes du pied prosodique (Figure1) s'effectue en calculant (en pourcentages) le rapport de la durée des syllabes St et Sə sur la durée totale du pied.

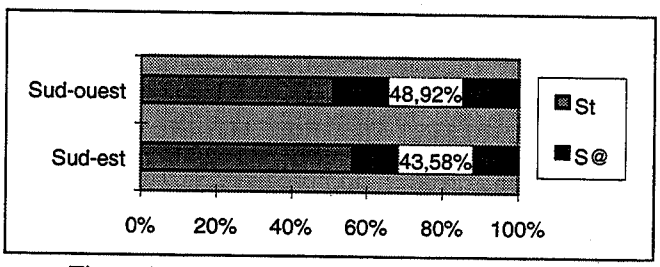


Figure 1 : Répartition des durées entre St et Sə

Bien que pour les deux groupes de locuteurs St soit en moyenne légèrement plus longue (53,63% de la durée totale du pied) que Sə (ici S@, en alphabet SAMPA), il arrive que cette tendance soit inversée, la durée de Sə n'excédant toutefois jamais 67% de la durée du pied. La durée relative de Sə est significativement plus importante pour les locuteurs Toulousains que pour les Marseillais : 48,92% contre 43,58% [F(1,219)=21,551 ; p<0,001]. L'analyse par locuteur fait apparaître que tous les locuteurs toulousains présentent des moyennes de Sə plus importantes que celle des marseillais, et qu'il n'y a pas de différence significative entre les locuteurs appartenant au même groupe régional, ce qui témoigne d'une grande homogénéité des résultats.

Il s'agissait ensuite d'observer si ces durées pouvaient être affectées par le contexte intonatif dans lequel se trouve le mot contenant un schwa. Celui-ci étant ici toujours le dernier mot d'une U.I. et par conséquent celui qui porte la marque de l'intonation. Trois modalités de base contextualisées ont été retenues : Assertion (A), Continuation (C) et Question (Q). Les résultats sont les suivants :

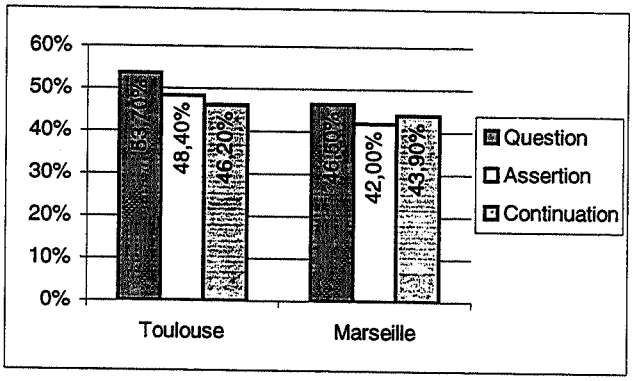


Figure 2 Moyennes Sə par modalité et origine

D'une manière générale, la modalité a une incidence sur la répartition des durées dans le pied prosodique [F (2,217)=6,797 ; p=0,0014]. En effet, la syllabe Sə est plus longue pour la question (50,1% pour les deux groupes) que pour les deux autres modalités qui présentent des moyennes similaires : Moyennes de 45,1% et 45,3% respectivement pour les continuations (C) et pour les Assertions (A). L'Anova donnant : Q/A, p=0,0011 ; Q/C, p=0,0010 et C/A, p=0,8579. L'effet lié a la modalité est

plus accusé pour les locuteurs Toulousains que pour les Marseillais. On relève également chez les locuteurs Toulousains une tendance à réaliser la syllabe S plus longue en contexte assertif qu'en contexte continuatif, mais cet effet ne permet pas de départager les deux groupes de locuteurs.

4.2 Répartition des durées dans les noyaux vocaliques

Il a été démontré expérimentalement (Astesano [Ast99]) pour le français général (non marqué dialectalement), que l'allongement d'une syllabe accentuée finale de groupe prosodique affecte essentiellement la rime de cette syllabe et donc son noyau vocalique. En supposant que ce phénomène s'applique également au français méridional, il est légitime de s'interroger sur ce qu'il advient de la distinction entre Toulousains et Marseillais que nous venons de mettre en évidence lorsque, à la différence de l'expérience précédente, on considère les noyaux vocaliques du pied à la place des syllabes, d'autant que les consonnes dites allongeantes du français ne remplissent pas leur fonction en méridional, notamment lorsqu'elles sont suivies de la voyelle / / (Carton et al [Car83]). Afin de répondre à cette question, nous avons calculé la durée cumulée de la voyelle tonique et du schwa subséquent et nous avons déterminé la durée relative (exprimée en pourcentages) de ces deux noyaux vocaliques par rapport à la durée cumulée. Nous avons également calculé la durée moyenne du schwa (en ms) pour les deux groupes de locuteurs (Tableau 1).

Tableau 1. Durées absolues et relatives du /E/ par région

	durée /E/ (ms)		pourcentage /E/	
	Toulouse	Marseille	Toulouse	Marseille
moyenne	93,17	73,42	46,30%	38,09%
nombre	116	106	116	106

L'examen du tableau montre que la durée moyenne du schwa est significativement plus importante pour les locuteurs Toulousains que pour les locuteurs Marseillais : 93,17 ms contre 73,42 ms [F(1,219)=18,454 ; p<0,0001]. D'autre part, les donnés relatives en pourcentages consignées dans le tableau font apparaître que la durée du schwa occupe 46,30% de la durée cumulée de la voyelle tonique et du schwa pour les locuteurs toulousains et 38,09% pour les locuteurs marseillais. [différence significative à p<0,0001 pour F(1,219)=25,276]. Il est donc établi que les résultats de cette expérience concernant la durée des noyaux vocaliques confirment pleinement la tendance observée précédemment où les durées des syllabes étaient prises en considération. Il apparaît également que cette tendance se trouve être amplifiée, dans la mesure où les différences entre les locuteurs des deux groupes se trouvent être encore plus accusées.

Dans les pieds prosodiques St-S dont nous étudions l'organisation temporelle, il existe une variable qui semble importante et qui concerne la durée intrinsèque du noyau vocalique de St. Nous savons en effet que les voyelles du français (Di Cristo [DiC80]) se répartissent en trois catégories de longueur en fonction de leur durée intrinsèque : Brève pour les voyelles hautes, Longue pour les voyelles basses et Très Longue pour les voyelles nasales. Compte tenu du fait que le facteur intrinsèque est la source d'une importante variation de la durée vocalique, on peut s'interroger à propos de son incidence éventuelle sur la durée du schwa dans le pied prosodique. Afin de répondre à cette question, nous avons calculé les durées moyennes du schwa et le pourcentage de la durée de cette voyelle par rapport à la durée cumulée des deux voyelles du pied, en prenant en considération la catégorie intrinsèque de la voyelle tonique (Tableau 2).

Tableau 2. Moyennes /E/ en fonction de la vt précédente

	moyennes /E/ (ms)			moyennes /E/ (%)		
Nature de la vt	Toul.	Mars	Total	Toul.	Mars.	Total
Brève	106,7	82,5	94,9	51,9%	42,6%	47,4%
Longue	92,8	72,7	83,5	48,1%	39,9%	44,3%
Très Longue	72,9	60,9	67,1	32,4%	27,4%	30,0%

Les données consignées dans le tableau 2 montrent à l'évidence que la durée intrinsèque de la voyelle tonique influence celle du schwa subséquent, dans la mesure où la durée de ce dernier varie en raison inverse de celle de la voyelle accentuée (colonnes de gauche). Il en va de même en ce qui concerne le pourcentage de la durée du schwa par rapport à la durée cumulée de la voyelle pénultième et finale. Il est intéressant d'observer que cet effet est significatif pour les données absolues [mesures en ms : F(1,216)=10,397 ; p<0,0001] et pour les données relatives [en % : F(1,216)=51,853 ; p<0,0001], ainsi que pour les deux groupes de locuteurs. Indépendamment de l'effet lié à la caractéristique régionale du locuteur, qui a été mis en évidence dans les expériences précédentes, nous relevons ici un effet généralisé de compensation de la durée des noyaux vocaliques au sein du pied prosodique. Ces résultats sont de nature à corroborer l'hypothèse selon laquelle le pied est une unité de gestion et de planification temporelle des éléments segmentaux qui le constituent.

5. Discussion et conclusion

Dans cette recherche expérimentale, qui se présente comme une contribution à l'analyse des caractéristiques prosodiques des parlers méridionaux de la France, nous nous sommes particulièrement intéressés à l'étude de l'organisation temporelle des pieds prosodiques St-S dans un parler régional du Sud-Est (Marseille) et du Sud-Ouest (Toulouse).

Les résultats de cette investigation ont permis de mettre en lumière un ensemble de faits que nous résumerons ainsi :

1. Pour les deux variétés régionales considérées, la syllabe tonique (St) est systématiquement plus longue que la syllabe subséquente comportant un schwa (S). Cette

hiérarchie temporelle au sein du pied s'applique d'une façon plus marquée en ce qui concerne les noyaux vocaliques respectifs de ces deux syllabes.

2. La durée absolue et la durée relative de la syllabe S dans le pied prosodique sont significativement plus longues pour les locuteurs Toulousains que pour les locuteurs Marseillais. À nouveau, cette tendance est plus marquée pour les noyaux vocaliques que pour les syllabes.

3. Le contexte intonatif (Assertion, Continuation, Question) exerce une influence sur la distribution des durées dans le pied pour les deux groupes de locuteurs, en ce sens que la durée de la syllabe S est plus longue avec une intonation de question que dans les autres contextes intonatifs.

4. Les caractéristiques intrinsèques de la voyelle tonique de St ont également une influence significative sur la durée du schwa, que cette dernière soit envisagée de façon absolue ou relative. La durée du schwa varie en raison inverse de la durée intrinsèque de la voyelle tonique dans les deux groupes de locuteurs considérés, sans que cet effet permette de les départager.

Ce travail a donc permis de mettre en évidence un ensemble de phénomènes qui confirment le bien fondé du choix du pied prosodique comme fenêtre d'observation de certaines caractéristiques susceptibles de distinguer les locuteurs Toulousains des locuteurs Marseillais. En effet, le point 2 sus-mentionné indique que la durée relative ou absolue du noyau vocalique ou de la syllabe S par rapport à St permet de distinguer les deux groupes de locuteurs. De plus, l'homogénéité des résultats analysés par locuteur suggère que ce phénomène pourrait être général, malgré le nombre limité de locuteurs.

Le point 4 suggère que le pied constitue bien un domaine de planification temporelle au sein duquel s'exercent des effets de compensation attestant de la cohésion de cette unité.

Nous savons par ailleurs que le pied final d'une Unité Intonative est le lieu de réalisation du contour mélodique caractéristique de la modalité de l'énoncé. Il serait intéressant de vérifier, à la lumière de nos résultats, si l'alignement des points clés constitutifs de ce contour avec les syllabes du pied s'effectue de la même manière - ou différemment - dans les deux variétés régionales qui nous intéressent. Ce point fera l'objet d'une étude ultérieure.

L'hypothèse initiale, selon laquelle le schwa ferait l'objet d'un traitement différencié par les locuteurs toulousains et les locuteurs marseillais, se trouve être pleinement validée, dans la mesure où l'analyse des données acoustiques et leur interprétation statistique permet d'expliquer les impressions auditives sur lesquelles s'est fondée cette hypothèse.

BIBLIOGRAPHIE

[Ast99] Astesano C. (1999) Rythme et discours : Invariance et sources de variabilité des phénomènes accentuels en français, Thèse de Doctorat. Université de Provence.

[Car83] Carton F., Rossi M., Autessere D. & Léon P. (1983) "Les accents du français", De bouche à oreille, Paris Hachette.

[Coq96] Coquillon A. (1996) Identification de l'accent méridional sur la base d'éléments prosodiques, Mémoire de Maîtrise. Université de Provence.

[DiC80] Di Cristo A. (1980) "Durée intrinsèque des voyelles du français", Travaux de l'Institut de Phonétique d'Aix n°7, pp. 211-235.

[Esp96] Espesser R. (1996) « Mes: un environnement de traitement du signal », Actes des XXI^{èmes} Journées d'Etudes sur la Parole, pp 447

[Hir98] Hirst D.J. & Di Cristo A (1998) Intonation systems : a survey of twenty languages, Cambridge University Press.

[Hir93] Hirst D.J. & Espesser R. (1993) "Automatic modelling of fundamental frequency using a quadratic spline function", Travaux de l'Institut de Phonétique d'Aix n°15, pp. 71-85.

[LeB96] Le Besnerais M. (1996) « Reconnaissance par des locuteurs monolingues et bilingues de l'espagnol et du français à partir de productions filtrées », Actes des XXI^{èmes} Journées d'Études sur la Parole, pp.47-50.

[Luc83] Lucci V. (1983) Étude phonétique du français contemporain à travers la variation situationnelle, thèse d'État, publication de l'Université de Grenoble, France

[Mai83] Maidment J.A. (1983) "Language recognition and prosody : further evidence", Speech, Hearing and Language 1, pp.131-141.

[Mor96] Mora Gallardo E. (1996) Caractérisation prosodique de la variation dialectale de l'espagnol parlé au Venezuela, Thèse de Doctorat. Université de Provence.

[Seg50] Seguy J. (1950) "Le français parlé à Toulouse", ed. Privat, 1978

[Sel77] Selkirk E.O. « The French foot : on the status of "mute" E », Colloquium on Current Issues in French Phonology, Indiana University

Des lexiques aux syllabes des langues du monde
Typologies et structures

Nathalie Vallée, Louis-Jean Boë, Ian Maddieson, Isabelle Rousset

Institut de la Communication Parlée, UMR CNRS 5009, Université Stendhal/INPG
BP 25 –38040 Grenoble cedex 9, France
Tél.: ++33 (0)476 82 41 19 - Fax: ++33 (0)476 82 43 35
Mél: vallee@icp.inpg.fr

ABSTRACT

This paper deals with the organization of sound structure in natural languages. As a first attempt to shed light on selection and restriction constraints in sound systems, we present typological results based on lexical and syllable structure of 13 ULSID languages. We claim that some of these tendencies can be analyzed in the substance-based linguistics framework.

1. INTRODUCTION

Les études récentes en phonétique-phonologie et en acquisition du langage montrent que l'étude de la syllabe est devenue un cadre nécessaire et incontournable pour comprendre le fonctionnement du langage et apporter des éléments intéressants de discussion à des questions et des problèmes d'ordre général non encore résolus : la syllabe est-elle l'unité fondamentale du langage humain ? Quels éléments pertinents pourraient appuyer l'existence phonétique, phonologique et psychologique de l'unité syllabique ? Dans quelles proportions le matériau phonologique universel des langues est-il basé sur des potentialités articulatoires reliables à la syllabe ? Comment éviter les hypothèses implicites sur l'organisation syllabique dans les modèles phonologiques ? Quels constituants déterminants de la syllabe faut-il analyser pour ambitionner une théorie générale de la syllabe ?

Toute unité lexicale, quelle que soit la langue, est fondée sur une concaténation d'un ou plusieurs types syllabiques existant dans cette langue. Une structure syllabique correspond à une concaténation de phonèmes décomposable traditionnellement en constituants C et V. Cependant, la nature de l'unité syllabique est encore actuellement l'objet d'un débat linguistique. Beaucoup on fait l'hypothèse que l'existence de l'unité syllabique ne pourrait être montrée que par l'existence de ses frontières. Des études pour délimiter la syllabe ont surtout été menées dans l'exploration des limites des phénomènes de coarticulation (cf. l'état de l'art de [Kra99]). Mais ces recherches, comme les études acoustiques d'ailleurs, n'ont pas livré de réponse complète et fiable sur l'unité syllabique. Depuis une dizaine d'années, les études sur la syllabe se penchent plutôt sur la recherche de caractéristiques articulatoires qui seraient plus spécifiques des attaques syllabiques et d'autres qui seraient des indices révélateurs de finales syllabiques. Ainsi, les études menées sur les asymétries phonologiques appuient le rôle organisationnel de l'élément syllabique dans les modèles phonologiques et vise à identifier la nature des représentations phonologiques de la syllabe pour enfin proposer une définition claire et précise de l'unité syllabique.

Pourtant, sans disposer d'une définition de la syllabe cohérente à plusieurs niveaux de description linguistique, la recherche et l'analyse des séquences généralement favorisées ou défavorisées dans les langues du monde, des fréquences des combinaisons inter- et intra-langue au niveau syllabique et lexical, des dépendances et indépendances distributionnelles des unités phoniques inter- et intra-syllabiques, montrent qu'il est tout à fait légitime de penser que l'organisation syllabique joue un rôle dans la structuration de la chaîne parlée et dans la formation des unités lexicales. Les tendances donnent de fortes raisons pour que l'organisation syllabique ne soit pas à considérer comme le fruit du hasard. Mais établir une typologie des langues basée sur la syllabe, en regroupant les structures syllabiques présentant des similitudes phoniques ou organisationnelles, est insuffisant pour élaborer des concepts phonologiques. Les typologies ne comportent pas en elles mêmes de pouvoir explicatif : elles aident à formuler des hypothèses qu'elles ne peuvent pas à elles seules prétendre vérifier. Aujourd'hui, il apparaît de plus en plus difficile, voire impossible, de tenter d'expliquer les tendances générales des structures sonores des langues du monde avec des principes internes de phonologie. Les explications des faits phonologiques sont actuellement recherchées dans les aspects articulatoires, acoustiques et psychologiques de la parole (cf. par exemple, [Lin 84] [Sch97] [Val99b]).

Les investigations que nous avons décidées de mener à l'ICP, dans le domaine de la syllabe, visent à prolonger nos travaux sur les contraintes dans l'organisation phonologique des structures sonores des langues naturelles. Dans l'esprit de ce qui est plaidé par [Lind84] à savoir, l'optimisation des séquences de segments à partir des propriétés des systèmes humains de production et de perception de parole, nous avons élaboré à l'ICP un modèle de prédiction des structures syllabiques CV, à partir d'une négociation entre *efficacité acoustique*, *coût articulatoire*, et *distinctivité systémique* [Ber95][Val99a].

Le critère méthodologique fondamental d'évaluation de nos modèles théoriques est que les résultats des simulations doivent être en forte adéquation avec un large échantillon de langues. Nous avons donc entrepris de mettre à jour des tendances dans l'organisation syllabique des langues du monde, non seulement afin d'améliorer et d'étendre les résultats des simulations à d'autres types de syllabes, mais aussi pour livrer un matériau intéressant qui permettra, à terme, de faire le point sur des discussions de questions générales de phonologie : peut-on établir une organisation universelle des structures syllabiques ? Peut-on dresser des inventaires de différents types de structures syllabiques selon les langues, selon les segments ? Dans quelle proportion les séquences associant consonnes et voyelles peuvent être déterminées par des caractéristiques articulatoires et perceptives des sons impliqués dans la cohorte ? Quelles sont les restrictions qui opèrent ? La différenciation des propriétés acoustiques et articulatoires des constituants de la syllabe est-elle déterminante dans l'organisation syllabique ? La distinctivité joue-t-elle un rôle dans l'émergence du lexique ?

Il ne s'agit pas, dans notre propos, d'analyser le statut linguistique de la syllabe, mais son contenu, afin de mettre en évidence les facteurs susceptibles de conditionner des choix ou des restrictions en faisant appel à la substance des unités sonores en contexte immédiat ou proches.

Dans l'état actuel des choses, nous nous limitons ici à un exposé de résultats préliminaires dans lequel figure un certains nombre d'observations qui permettent déjà de dégager des caractéristiques typologiques basées sur la syllabe et de préciser des considérations générales sur les types syllabiques de 13 langues.

2. ULSID

Grâce à Ian Maddieson, nous disposons d'une base de données contenant les lexiques de 32 langues découpés en syllabes [Mad92]. Baptisée ULSID, acronyme de UCLA - *Université de*

Los Angeles - Lexical and Syllabic Inventory Database, elle contient des langues sélectionnées dans un soucis de diversité à la fois génétique et géographique : hétérogénéité des familles de langues et très large répartition géographique. Les langues retenues disposent toutes d'un dictionnaire ou d'un lexique dont les entrées sont soit phonétiques, soit phonologiques, soit orthographiques lorsque le code graphique de la langue est aisément interprétable avec un code phonétique. La totalité des lexiques est alors transcrite en ASCII avec deux types de séparateurs : l'un indiquant le découpage de chaque entrée en syllabe, l'autre codant la séparation entre attaque, noyau et coda de chaque syllabe. Nous donnons à titre d'exemple, trois entrées en ASCII du lexique du kwakw'ala (langue almosan), 2 des items étant dissyllabiques, un troisième trisyllabique :

k u Xw s .? a
k u Xw .ts' a .n a
k u Xw .? i d

Lors de l'implantation d'ULSID à l'ICP, nous avons préféré harmoniser la transcription entre les lexiques, en adoptant les symboles conventionnels de l'API (1996) et en conservant les informations sur le découpage. Actuellement, 13 langues sur 32 ont été transcodées en retournant aux sources phonétiques et phonologiques de chaque langue. Notons la diversité dans la taille et l'inventaire des systèmes vocaliques et consonantiques de ces langues. C'est de cet échantillon (cf. Table 1), resté représentatif de la diversité mentionnée ci-dessus, que nous avons tiré les premiers résultats.

La taille des lexiques est portée sur la figure 1. Au total, nous disposons d'un peu plus de 160 500 syllabes résultant du découpage de 60 000 entrées lexicales, avec une moyenne de 4 560 termes par langue.

Table 1: Les 13 langues tirées d'ULSID

langue	famille, branche	lieu
wa	austro-asiatique, palaung	Chine
kannada	dravidien, sud-dravidien	Inde : état du Mysore
sora	austro-asiatique, mounda	Inde: états d'Orissa et d'Assam
thaï	kam-taï, kadaï,	Thaïlande, Birmanie, province du Yunnan
nyahkur	austro-asiatique, môn	Thaïlande (centre)
ngizim	afro-asiatique, tchadique	Nord-est du Nigéria
afar	afro-asiatique, couchitique	corne orientale de l'Afrique
kanouri	nilo-saharien, saharien	Nigéria, Niger, Tchad, Cameroun
navaho	na-déné, athabaskan	USA : Arizona, Colorado, Nouveau-Mexique, Utah
kwakw'ala	Amérique du Nord, almosan,	Canada : Colombie Britannique
quechua	Amérique du Sud, Andin	Pérou, Équateur, Bolivie, Colombie
yup'ik	eskimo-aléoute, eskimo	Alaska
finnois	ouralo-altaïque, finno-ougrien	Finlande

Les emprunts récents, dont les référents sont d'ordre technologique, politique ou culturel, ont été soustraits des lexiques par [Mad92]. Les informations concernant le découpage syllabique ont été livrées soit par les dictionnaires, soit à partir d'informateurs, locuteurs natifs des langues concernées.

3. STRUCTURE SYLLABIQUE DES UNITES LEXICALES ET TYPOLOGIE DES LANGUES

Ne disposant pas de données sur la nature des unités lexicales, elles sont toutes prises en compte dans les résultats.

La répartition des langues en fonction du nombre de syllabes contenues dans les entrées lexicales, permet de dégager 4 types :

- Le type 1 regroupe les langues qui présentent une distribution des unités lexicales telle que 40% au moins d'entre elles sont monosyllabiques, entre 20 et 40% sont dissyllabiques, entre 10 et 20% sont trisyllabiques (navaho, thaï) ;
- Le type 2 rassemble les langues totalement ou très majoritairement monosyllabiques : le wa (100%), le nyahkur (70% d'unités monosyllabiques et 30% de dissyllabes) ;
- Sous le type 3 (Figure 2), se classent les langues qui présentent une majorité d'items dissyllabiques, avec moins de 10% d'unités monosyllabiques et environ un tiers d'entrées lexicales trisyllabiques ;
- Le type 4 regroupe le finnois, le kanouri et le yup'ik, qui ne présentent que très peu d'unités monosyllabiques dans leur lexique, mais une majorité d'unités trisyllabiques, et 20 à 40 % de dissyllabes, 25% de quadrisyllabes, de 2 (kanouri) à 12 % (finnois) de quinquasyllabes.

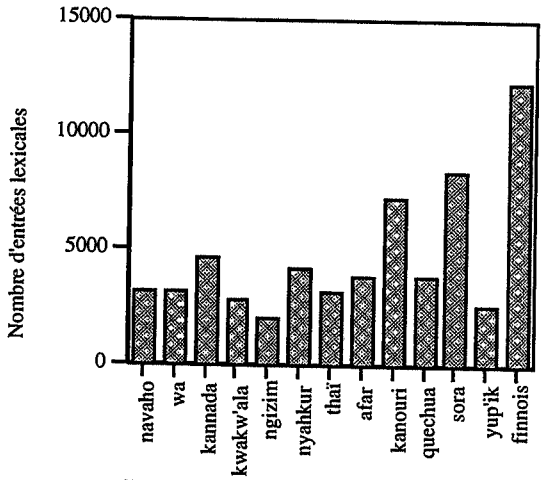


Figure 1 : Taille des lexiques

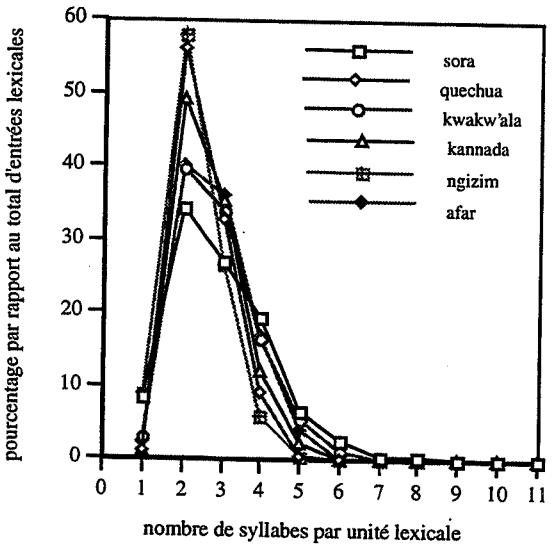


Figure 2 : Répartition des unités lexicales par nombre de syllabes, la plus représentée dans les 13 langues d'ULSID

De toute évidence, ce sont les structures lexicales di- et trisyllabiques qui sont très largement favorisées dans les lexiques des langues de notre échantillon, mises à part, bien entendu, celles relevant du type 2. Toutefois, 8 langues sur 13 favorisent les dissyllabes, contre 3 privilégiant les unités trisyllabiques (Figure 3). Notre typologie montre que les éléments lexicaux de plus de 5 syllabes sont marginaux (entre 0 et moins de 4% du lexique des langues), ceux à 5 syllabes constituant au maximum

un peu plus de 10 % du lexique pour le finnois et le yup'ik (de type 4) et entre 0 et moins de 6% du lexique pour les autres langues.

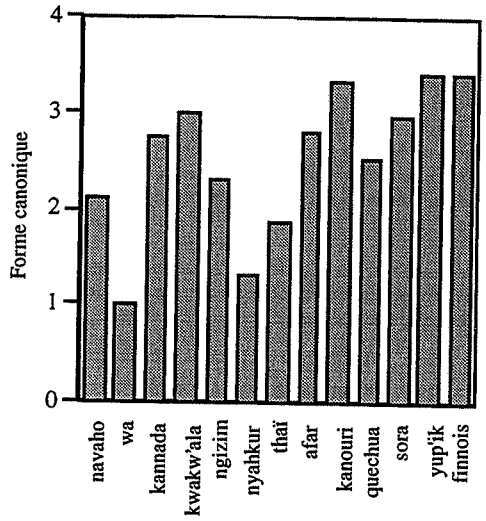


Figure 3 : Forme canonique (Fc) des langues selon la formulation de [Mon72] : Fc correspond au rapport entre le nombre total de syllabes pour l'ensemble d'un lexique donné et le nombre d'entrées lexicales. Les valeurs obtenues pour Fc montrent bien la préférence des langues pour les formes lexicales di- et trisyllabiques.

4. RENDEMENT SYLLABIQUE

Pour les lexiques de moins de 5000 termes, on ne relève pas de corrélation très nette entre la taille du lexique et le nombre total de syllabes qui le constituent. Seules trois langues (sora, kanouri, finnois) présentent entre 7 000 et 12 000 entrées lexicales et un nombre de syllabes allant de 19 700 à plus de 41 000. Pour ces 3 langues, on note un large nombre de syllabes lié à la taille plus importante du lexique.

Sans faire d'équivalence typologique sur la nature des segments qui composent les syllabes, nous avons procédé à un regroupement des syllabes identiques pour chacune des langues. Nous avons déterminé le rendement syllabique comme étant le rapport entre le nombre total de syllabes obtenues dans le découpage d'un lexique donné et le nombre de syllabes différentes comptabilisées après regroupement (Figure 4).

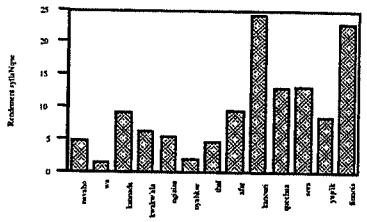


Figure 4 : Rendement syllabique pour chaque langue, calculé comme le rapport du nombre total de syllabes et du nombre de syllabes différentes après regroupement des syllabes identiques.

On constate un rendement élevé pour le finnois ou le kanouri qui sont des langues relevant du type 4, et qui se caractérisent donc par une forte proportion d'unités lexicales trisyllabiques. Les rendements les plus faibles sont rencontrées pour les langues du type 2, c'est-à-dire, celles totalement ou très majoritairement monosyllabiques. Un rendement proche de 1 signifie que la fréquence des syllabes dans le lexique de la langue concernée est relativement faible. Un rendement proche de 5 caractérise les langues du type 1 (navaho et thaï), alors que les langues du type dominant (type 3) ont un rendement syllabique allant de plus de 5 à 14.

5. TYPOLOGIE ET TENDANCES DES STRUCTURES SYLLABIQUES

La décomposition des syllabes de chaque lexique en constituants C et V, et leur regroupement en structure identique, révèlent que le nombre de type est relativement restreint, quelle que soit la langue, et qu'il varie dans un intervalle allant de 4 à 11, avec une moyenne de 7.7 types par langue. Le contingent de types de syllabes pour chaque langue est totalement indépendant de la taille du lexique, du nombre total de syllabe et du rendement : citons le kannada, le nyahkur et le kanouri qui possèdent chacune 7 types de structures syllabiques avec respectivement 4 559, 4 188, 7 211 entrées lexicales, et 12 126, 5 503, 19 773 syllabes au total, ainsi que des rendements de 9.12 pour le kannada, 1.88 pour le nyahkur et 24.23 pour le kanouri. Cette typologie de la structure syllabique appuie l'existence de restrictions qui organisent la syllabe dans les langues naturelles : les 160 517 syllabes, totalisées sur l'ensemble de notre corpus, se répartissent dans 16 types de structures syllabiques (Table 2).

Table 2: Liste et occurrence des structures syllabiques rencontrées dans les 13 langues d'ULSID

type	0/000	type	0/000
CV	5448	CCCVC	2
CVC	3615	C	1.4
V	440	CVCCC	1.4
VC	247	CCCV	0.3
CCVC	126	VCCC	0.3
CVCC	63	CVCCCC	0.2
CCV	50	CCVCC	0.2
VCC	5	CC	0.2

Les syllabes fermées présentent plus de diversité : 11 types contre 5 pour les syllabes ouvertes. Les deux structures que l'on rencontre dans toutes les langues de la base sont CV et CVC, mais pour 10 d'entre elles, c'est la première qui est la plus répandue, le wa, le nyahkur et le thaï possédant de préférence des syllabes CVC. De manière plus générale, ces trois langues favorisent, contrairement aux autres, les syllabes fermées (Figures 5 & 6). On note également que 8 langues sur 10 comptent 60% ou plus de syllabes dans le type CV. De ce fait, la tendance très forte qui se dégage de cette typologie est que les groupements consonantiques intra-syllabiques sont nettement défavorisés : on ne les rencontre que dans 1.26% des quelques 160 500 syllabes de notre corpus. Si on se penche sur la place de ces groupements consonantiques dans la syllabe, nous constatons que la plupart (67%) occupent la position initiale, contre 33% en finale de syllabe. Ce résultat est renforcé par le fait que les consonnes complexes (celles qui superposent à une articulation de base des modes articulatoires tels que la labialisation, l'aspiration, la glottalisation, la palatalisation, la prénasalisation... mais qui n'occupe qu'une position C dans notre corpus) sont bien plus fréquentes en attaque de syllabe. Encore reste-t-il à déterminer s'il existe des tendances dans la nature des consonnes qui apparaissent dans les groupements.

Les articulations vocaliques ou consonantiques élaborées peuvent être rencontrées parmi les syllabes les plus fréquentes dans une langue donnée (exemples [k a: n] [tõi] ou [kõ w a: m] qui sont respectivement les syllabes de rang 2, 4 et 5 du thaï). La comparaison des 15 syllabes les plus fréquentes pour chaque langue, montre qu'elles possèdent majoritairement une attaque constituée de plosives sourdes, vélaires et coronales [k t], devant les latérales coronales [l], les nasales coronales et bilabiales [m n] et la fricative coronale sourde [s]. Les noyaux de ces syllabes fréquentes sont pour plus d'une syllabe sur deux occupés par la voyelle ouverte [a], devant [i] et [u]. Nous remarquons également que [a] constitue le noyau le plus répandu des syllabes

qui présentent une attaque et une coda vide (type V). Lorsqu'une langue n'est pas en adéquation avec cette tendance, c'est que la structure V est marginale ou inexistante sur l'ensemble du lexique (navaho, wa, nyakhur, thaï). Il est plus difficile de tirer une observation générale sur la nature des consonnes en coda de syllabes fréquentes dans une langue car pour beaucoup c'est la structure CV qui prédomine. À noter une forte proportion de [n] dans cette position. Parmi les syllabes les plus répandues dans une langue donnée, l'inventaire des consonnes en coda semble bien plus restreint que l'inventaire des possibilités pour les consonnes en attaque.

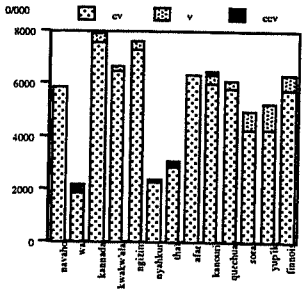


Figure 5 : Répartition des types de syllabes ouvertes par langue (seuls figurent les plus fréquents).

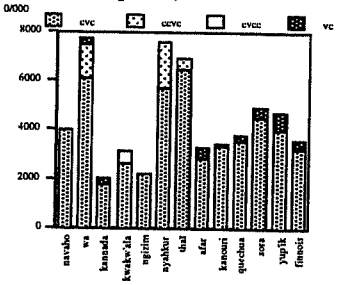


Figure 6 : Répartition des types de syllabes fermées par langue (seuls sont considérés les plus répandus).

6. DISCUSSION

Si la syllabe est l'unité de base de la parole, on doit pouvoir trouver les traces d'une organisation syllabique dans les langues du monde. Il s'agit pour nous de récupérer, par l'observation des structures syllabiques des lexiques de langues naturelles (à la différence de [Jan86] qui a considéré la fréquence des syllabes dans des textes de 4 langues), les informations nécessaires qui nous permettront d'expliquer et simuler, grâce à la modélisation, l'émergence syllabique et les formes prises par les syllabes. Nos investigations sont pour l'heure largement insuffisantes. N'ont pas encore été étudiées la fréquence des syllabes sur l'ensemble du corpus, les contraintes sur les cohortes, les associations consonantiques et consonne-voyelle en fonction de la position dans la syllabe (attaque, coda), ainsi que la nature de la voyelle en fonction du lieu d'articulation de la consonne et ceci, en nous appuyant sur des résultats typologiques que nous avons déjà acquis [Val99b].

Nous faisons l'hypothèse que la fréquence importante de la structure CV, la marginalisation des groupements consonantiques intra-syllabique et la forte proportion d'unités lexicales dissyllabiques dans les langues pourrait très bien être liée, non pas à une exigence formelle, mais à des contraintes de production. Notre objectif est de tenter d'analyser les points cités ci-dessus en considérant le geste articulatoire de fermeture et d'ouverture du tractus vocal rythmée par la mâchoire, selon la théorie *frame/content* de [Mac98]. Dégager les schèmes structurels des différents lexiques est une étape qui s'imposera dans cette recherche, de même que confronter les tendances des structures syllabiques aux données de l'ontogenèse [Vih96] puisque les cohortes CV sont aussi les syllabes canoniques du babillage de l'enfant, quel que soit son environnement linguistique [Loc83]. Ces tendances peuvent aussi étayer l'existence d'un attracteur CV au détriment des autres types de syllabe, et de VC en particulier, qui fait converger les gestes syllabiques vers une forme plus stable et mieux résistante aux contraintes de production. Cet attracteur est mis en évidence avec un paradigme de contrainte de production basé sur le débit, le locuteur devant produire des syllabes avec un débit de plus en plus accéléré, synchronisé sur un métronome [Tul91]. Désorganisant le contrôle de synchronisation supposé coûteux, les syllabes VC "switche" vers l'attracteur CV. Par ailleurs, différentes études [Kra99] mettent en évidence qu'en attaque de syllabe, la force articulatoire est plus grande, le geste plus précis, et que les segments montrent une plus grande stabilité, une résistance plus importante à la coarticulation. De tels avantages expliqueraient que les langues favorisent cette structure.

À terme nous voulons pouvoir trancher sur la question des "hasards" du lexique en montrant que les syllabes ne répondent pas seulement à un pur formalisme, mais aussi à des contraintes de production et de perception de parole. Serait du même coup remis en cause le principe d'arbitrarité des unités linguistiques hérité du structuralisme.

BIBLIOGRAPHIE

[Ber95] Berrah A.R., Boë L.J., Schwartz J.L. (1995) Emergent Syllable using Articulatory Acoustic Principles. *XIIIth ICPhS*, 1, 396-399.

[Jan86] Janson T. (1986) Cross-linguistic Trends in the Frequency of CV Sequences. *Phonology Yearbook* 3, 179-195.

[Kra99] Krakow R. (1999) Physiological organization of syllables : a review. *J. of Phonetics 27*, 23-54.

[Lin84] Lindblom B., MacNeilage P.F., Studdert-Kennedy, M. (1984) Self-organizing Pocesses and the Explanation of Languages Universals. *Explanation of Language Universals*, Butterworth B., Comrie B. & Dahl O., Mouton, 181-203.

[Loc83] Locke J.L. (1983) *Phonological acquisition and change*. Academic Press, New-York.

[Mac98] MacNeilage P.F. (1998) The Frame/Content theory of evolution of speech production, *Behavioral and Brain Sciences*, 21,4, 499-511.

[Mad92] Maddieson I. (1992) The structure of segment sequences, UCLA Working Papers in Phonetics 83, 1-8.

[Mon72] Monino Y., Roulon P. (1972) Phonologie du Gbaya Kara 'Bodoe, *SELAF 31*, Paris.

[Sch97] Schwartz J.L., Boë L.J., Vallée N., Abry C. (1997) The Dispersion-Focalization Theory of Vowel Systems. *J. of Phonetics* 25, 255-286.

[Tul91] Tuller B., Kelso J.A.S. (1991) The Production and Perception of Syllable Structure. *JSHR* 34, 3, 501-508.

[Val99a] Vallée N., Berrah R., Boë L.J., Schwartz J.L. (1999) Syllabes CV : modèle d'émergence et simulation. *Journées d'Études Linguistiques " SyllabeS"*, AAI, Nantes, 128-135.

[Val99b] Vallée N., Boë L.J., Stefanuto M. (1999) Typologies phonologiques et tendances universelles. Approche substantialiste. *Linx*, 31-54, numéro spécial 1999.

[Vih96] Vihman M.M. (1996) *Phonological development: The origin of language in the child.*: Blackwell Publishers Ltd.

Nos remerciements à Delphine Zouyed, Céline Bard pour leur contribution à l'implantation d'ULSID. Une partie de cette recherche est financée par le GDR-CNRS 1954 "Phonologie".

Synthèse

Un modèle neuronal pour la prédiction de la durée des syllabes de la langue arabe

A.Chehab[(1)], *A.Zaki* [(2)], *A.Rajouani* [(3)]

[(1)] READ, LORIA, Campus Scientifique BP 239 - F54506 Vandœuvre-lès-Nancy, France
[(2)]Equipe Signal et Image, ENSERB. B.P 99, F-33 402 TALENCE Cedex, France
[(3)]LEESA, Faculté des Sciences. B.P 1014 - Rabat, Maroc
Tél.: ++33 (0)383 59 20 00 - Fax: ++33 (0)383 27 83 19
Mél:Ahmed.Chehab@loria.fr, zaki@goelette.tsi.u-bordeaux.fr

ABSTRACT

This paper present a neural-network-based model of syllable duration which is developed for amelioration of the naturalness of Arabic TTS. Given a set of factors influencing the duration of syllable a neural network is used to predict the syllable duration, different mappings of theses factors to values suitable for networks with binary and analog input nodes have been applied. The highest correlation coefficient between the observed and predicted syllable duration of test set is 0.852.

1. INTRODUCTION

L'amélioration du naturel d'un système de synthèse de la parole à partir du texte, suppose l'estimation exacte des paramètres prosodiques (durée des segments, Fréquence fondamentale F0, intensité).

La durée des segments phonétiques est infectée par un ensemble de facteurs phonologiques , morphologiques, syntaxiques [Kla79], qui interagissent de manière complexe, et rendent la prévision de la variation de la durée difficile.

La présentation de ces facteurs comme un vecteur d'entrée d'un réseau multicouche à connexion totale basé sur la rétropropagation d'erreur, permet de prédire la durée relative. La représentation du vecteur d'entrée (facteurs) et celle de sortie (durée à prédire), influent sur la performance du modèle.

2. LE MODELE

Le but de notre modèle est la détermination de la durée des unités syllabiques de la langue arabe, à partir de la connaissance des paramètres ou facteurs qui décrivent cette syllabe.

Dans ce qui suit on présente une descriptions des différents facteurs utilisés comme informations nécessaires pour la prédiction de la durée.

- **Type de la syllabe (a)** : le nombre de syllabes de la langue arabe est cinq [Raj85], répartis en syllabes ouvertes ou fermées, et en syllabes courtes ou longues, CV, CVV, CVC, CVVC, CVCC.

En morphologie non linéaire ou métrique la syllabe connaît une répartition en Attaque et Rime, la Rime se subdivise en noyau et coda [Bre95].

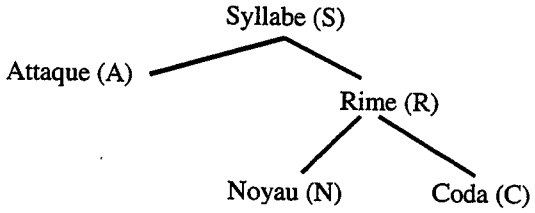


Figure1: La décomposition morphologique de la syllabe en Attaque, Rime, Coda et Noyau.

- **Type de la consonne (b)** : six types à considérer, on les distingue selon les trois facteurs suivants:
  - Vibration des cordes vocales,
  - Franchissement de l'air à travers le conduit vocal,
  - Le mode d'articulation.

Une consonne, elle est soit fricative, Occlusive, Liquide, nasale, Vibrante, ou semi-voyelle.

- **Nature de la voyelle (c)** : les voyelles de la langue arabe sont en nombre de six; trois voyelles courtes, et trois voyelles longues. /a/,/i/, /u/, /aa/, /ii/, /uu/.

- **Position de la syllabe dans le mot (d)** :quatre niveaux à considérer, position initiale, finale, médiane, ou monosyllabique, s'il s'agit d'un mot à une syllabe.

- **L'accent (e) :** les règles régissant la place de l'accent dans un mot sont définies d'après [Raj87], comme suit :

La position et la distribution de l'accent dépend du nombre et du type de syllabes contenu dans le mot. Les monosyllabiques reçoivent un accent primaire, les disyllabiques reçoivent un accent primaire et un accent de troisième niveau, alors que les polysyllabiques reçoivent en plus de l'accent primaire, un accent secondaire et un accent de troisième niveau.

Si toutes les syllabes du mot sont de type CV alors c'est la première syllabe du mot qui porte l'accent primaire, les autres syllabes reçoivent un accent de troisième niveau.

Exemple : /daxala/

Si dans le mot, il y a une seule syllabe longue, alors cette syllabe reçoit l'accent primaire, les autres auront un accent de troisième niveau.

Exemple : /kaafaha/

Si le mot contient deux syllabes longues ou plus, alors c'est la syllabe la plus proche de la fin du mot (la dernière syllabe n'est pas comptée) qui reçoit l'accent primaire, la syllabe longue la plus proche du début du mot reçoit un accent secondaire, les autres syllabes reçoivent un accent de troisième niveau.

Exemple : /mus[alahaatun/

L'influence de l'accent s'étend sur plus qu'une syllabe, pour une syllabe donnée on tient compte de l'accentuation de la syllabe suivante et précédente.

- **Nature de la consonne (f) :** les consonnes de la langue arabe sont classées selon le voisement, la gémination et l'emphatique. Pour la gémination, on considère trois niveaux: gémination avec signe de gémination, gémination sans signe de gémination (succession de deux segments de la même consonne) exp: /^an nulaahida/ ou une consonne non géminée.

- **Nature du mot (g) :** les mots sont ou grammaticaux, ou lexicaux, en se basant sur la prévision du mot.

Deux autres facteurs à considérer :

- La nature de la consonne nasale, soit c'est une vraie consonne soit elle dérive d'une voyelle (les signes de voyellation).
- La nature de la liquide: soit elle dérive d'un mot définit, elle est donc toujours au début du mot, ou non.

3. LE RESEAU NEURONAL

La figure 2 montre la structure d'un modèle de prédiction de la durée syllabique de la langue arabe.

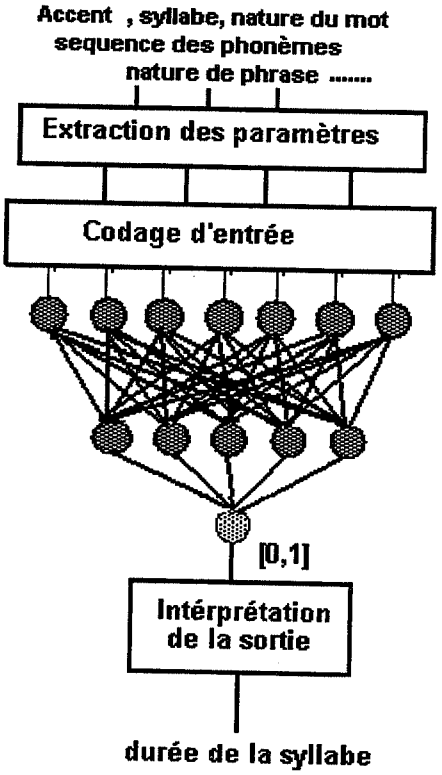


Figure 2 : la structure du réseau multicouche appliquée pour la prédiction de la durée des syllabes

En associant une valeur à chaque facteur (décrit précédemment) , la durée des syllabes peut être prédite par le réseau.

Notre modèle est un réseau multicouche qui se base sur l'algorithme de rétropropagation d'erreur décrit par Lippmann [Lip87] .

Parmi les paramètres du réseau qui varient il y a le nombre de couches cachées, le nombre de nœuds dans chaque couche, les valeurs initiales . Ces paramètres ont une influence sur la performance du modèle. La durée des syllabes, et les facteurs d'entrée doivent être codés, le type du codage affecte la performance du modèle.

Les nœuds de chaque couche de notre modèle ont une fonction sigmoïde ($1/1+e^{-u}$) comme fonction d'activation

Les facteurs qui influent sur la durée sont les entrées du réseau; un choix évident est de les représenter par des entrées binaires ou analogiques. Une combinaison entre les entrées binaires et analogiques reste aussi une possibilité mais n'a pas été traiter dans notre présent travail.

De meilleurs résultats sont obtenus avec la représentation suivante des entrées/sorties.

3-1 Entrées binaires

Chaque entrée correspond à l'état d'un facteur, elle prend l'une des valeurs 0 ou 1 selon que ce facteur contribue ou non dans la formation de la durée.

3.2 Entrées analogiques

Chaque facteur des N, représente une entrée qui prend l'une des valeurs 0 , 1/n-1, 2/n-2,.....n-1/n-1 . Pour chaque facteur on calcule la moyenne sur la base de durée des syllabes dont ce facteur apparaît. Le facteur dont la moyenne correspondante est forte aura comme valeur d'entrée la plus grande, et celui qui à la moyenne la plus faible aura comme valeur d'entrée 0. Et de même pour les autres.

3.3 Sortie analogique

La durée des syllabes observées varie de façon continue entre 60ms et 350ms, et afin de l'adapter à la sortie du réseau trois représentations ont été appliqués dans le but d'avoir une valeur de sortie comprise entre 0 et 1. Les trois fonctions utilisées sont :

$$- S(i) = a * S\_r(i) + b.$$

$$- S(i) = {S\_r(i)}/{Dur\_Max}$$

$$- S(i) = a * Log(S\_r(i)) + b.$$

- S : est la sortie normalisée.
- S_r :La durée observé
- *Dur_Max* : la durée Maximal Observé.
- a , b sont des coefficient à calculer.

Seule la fonction logarithmique qui à été utilisé dans le modèle final.

4. L'APPRENTISSAGE DU RESEAU NEURONALE

L'algorithme de rétropropagation d'erreur décrit par Lippmann [Lip87], est appliqué au cours de l'apprentissage du réseau.

Une étape très importante dans la réalisation du modèle de contrôle de la durée, est la préparation d'un nombre suffisant des exemples pour l'apprentissage et le test.

Un corpus de six paragraphes, environ 40 phrases, lu par un seul locuteur, est manuellement segmenté à l'aide du spectrogramme et de la perception, il en résulte une base de données de 1950 syllabes .

Pour la prédiction de la durée des syllabes nous considérons 15 facteurs, il en résulte près de $4{,}8.10^7$ combinaisons, non tous réalisables.

On à utilisé 85% de la base des syllabes pour l'apprentissage , et 15% pour le test.

Lors d'entraînement du réseau le coefficient de corrélation est calculé dans le but d'utiliser le modèle le plus performant. La performance du modèle dépend d'un certain nombre de paramètres qui décrivent le réseau, à part le type d'entrée et sortie déjà mentionnés dans la section 3, les paramètres suivants ont été modifiés durant les entraînements.

4.1 Nombre et dimension des couches cachées

On a utilisé des réseaux à une seule couche cachée, et de deux couches cachées, les meilleurs résultats ont été remarqués lors de l'utilisation d'un réseau à deux couches cachées.

Le nombre de nœuds dans chaque couche cachée est modifié au cours des entraînements entre 4 et 150.

La tableau ci-dessous montre la performance d'un modèle à deux couche cachés par rapport au modèle qui utilise une seule couche caché.

Table 1, Coefficient de corrélation en fonction du nombre d'itérations pour une (coef_1) et deux (coef_2) couches cachées, dans l'apprentissage et le test.

Nbr d'itérations	Coef_1 *app*	Coef_1 *test*	Coef_2 *app*	Coef_2 *test*
10^2	0,6172	0,5303	0,7522	0,7143
10^3	0,6755	0,6015	0,8157	0,7742
10^5	0,7011	0,6505	0,8464	0,8145
2.10^5	0,7304	0,6901	0,8715	0,8352
5.10^5	0,7615	0,7294	0,8994	0,8444
7.10^5	0,7802	0,7305	0,9105	0,8495
10^6	0,7978	0,7449	0,9288	0,852

4.2 Nombre d'itérations

Le grand nombre d'itérations dépend du pas d'algorithme de rétropropagation d'erreur utilisé, un mauvais choix de pas conduit à la divergence de l'algorithme.

Nous avons utilisé 64 nœuds d'entrées binaires ou 15 nœuds d'entrées analogiques, et un nœud dans la couche de sortie.

5. RESULTATS

Les différents réseaux testés sont comparés entre eux en calculant le coefficient de corrélation entre les durées observées et celles prédites par le réseau.
Le plus grand coefficient de corrélation est observé pour des entrées binaires. La figure 3 montre l'évolution du

coefficient de corrélation en fonction du nombre d'itérations, les autres paramètres sont optimaux.

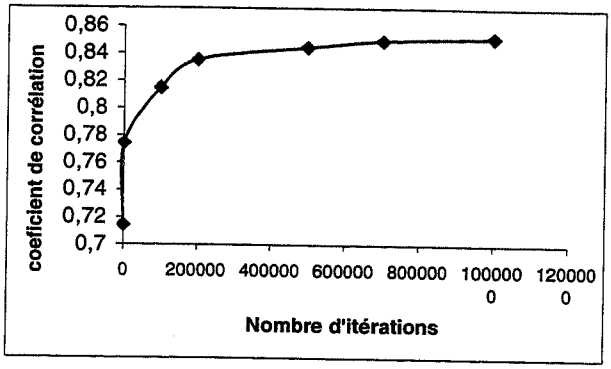


Figure 3 : la variation du coefficient de corrélation lors du test en fonction du nombre d'itérations

Et pour tester la contribution de chaque facteur dans la construction de la durée, la méthode du variant simple proposé par [Cam90] à été adopté dans ce but.

En effet, alternativement on met les valeurs des nœuds d'entrées relatives à chaque facteur à zéro et on calcule le coefficient de corrélation Coef-test, plus le coefficient de corrélation est petit plus la contribution est grande.

Le tableau suivant donne les valeurs des coefficients de corrélation en l'absence de chaque facteur, et le pourcentage de contribution de chaque facteur dans la construction de la durée pour un coefficient de corrélation de 0.852.

Table 2: Coefficient de corrélation et pourcentage de contribution de chaque facteur dans la construction de la durée de la syllabe.

Facteur	Signe	Coef-test
Type de la consonne de l'attaque	b_1	0.3974
Position de la syllabe dans le mot	d	0.5665
Voisement de la consonne précédente	f_{-1}^{v}	0.5763
l'accent de la syllabe	e	0.6032
l'emphase de la consonne de l'attaque	f_0^{e}	0.6224
Nature du mot	g	0.6516
Voisement de la consonne de la coda	f_1^{v}	0.6529
Type de la syllabe suivante	a_{+1}	0.6561
Gémination de la consonne précédente	f_{-1}^{g}	0.6597
Voisement de la consonne de l'attaque	f_0^{v}	0.6896
Gémination de la consonne suivante	f_{+1}^{g}	0.6992
Type de la consonne précédente	b_{-1}	0.7277
Nature de la voyelle de la syllabe	c	0.7411
Type de la syllabe précédente	a_{-1}	0.7582
Gémination de la consonne de l'attaque	f_0^{g}	0.7804

6. CONCLUSION ET PERCPECTIVES

Dans ce travail nous avons présenté un exemple de modélisation non-paramétrique de la durée syllabique en faisant appel au réseaux de neurones. La simulation faite a permet d'aboutir une estimation acceptable des durées syllabiques. La prochaine étape consiste à appliquer le modèle sur un corpus d'entraînement assez large et qui englobe les différentes réalisations syllabiques de l'arabe. En effet, les recherches qui ont été faites sur la prosodie de la langue arabe [Raj87], on montrées l'existence d'une relation étroite entre la prosodie et l'accent qui se base sur la structure syllabique. Et dans ce sens ce modèle et conçu pour qu'il soit intégré avec un modèle intonatif qui est en cours de développement [Zak00] au niveau du système de synthèse de la parole arabe à partir du texte développé au LEESA.

BIBLIOGRAPHIE

[Kla79] Klatt D. "Synthesis by Rule of Segmental Durations in English Sentences". In Frontiers of Speech Communication Research, edited by B. Lindblom and S. Öhman (Academic, London), pp. 287-301, 1979.

[Raj85] Rajouani A. et al, "Etude de la Gémination des Occlusives en arabe ", 15ème J.E.P, pp. : 16-18. 1985

[Raj87] Rajouani A. et al, " Syntheses by Rule of Arabic language", Proc of European conf on Speech Tech, vol 1 pp : 29-32 Edinburg, 1987.

[Lip87] Lippmann R. P. "An Introduction to Computing With Neural Nets" IEE ASSP Magazine, Vol 4, pp 4-22, april 1987.

[Raj87] Rajouani A. et Al., Synthèse et Perception de l'Accent Lexical en arabe. Actes GALF, 16ème JEP, pp. 302-305, Hammamet, 1987.

[Cam90] Campbell W. " Analog I/O nets for Syllable Timing", in speech communication, vol 9, pp 57-61, North-Holland, 1990.

[Bre95] Breen.A.P. "A Simple Method of Predicting the duration of Syllables" EUROSPEECH'95 September 1995 pp:595-598.

[Zak00] Zaki A. et Al "Contours intonatifs de la phrase interrogative en arabe", accepté au XXIIIèmes JEP, Aussois, 2000.

Le projet EULER : la synthèse de parole générique multilingue

Michel BAGEIN, Thierry DUTOIT, Fabrice MALFRERE, Vincent PAGEL, Alain RUELLE, Nawfal TOUNSI, Dominique WYNSBERGHE

Faculté Polytechnique de Mons, TCTSLAB, Avenue Copernic Bâtiment Multitel Materia Nova
7000 MONS (BELGIQUE)
Tél.: ++32 65 37 47 36 - Fax: ++32 65 37 47 29
Mél : euler@tcts.fpms.ac.be - [bagein,dutoit,pagel,tounsi,ruelle,wynsberg]@tcts.fpms.ac.be – malfrere@babeltech.com
Web : http://www.tcts.fpms.ac.be - http://tcts.fpms.ac.be/synthesis/euler/

RESUME

EULER est un projet de recherche et de développement mis en place par le groupe de recherche en Synthèse de la Parole de la Faculté Polytechnique de Mons. Ce projet intègre progressivement les résultats d'autres projets de recherche tant en synthèse de parole qu'en traitement du langage naturel. L'objectif de ce projet est de réunir, grâce à une collaboration internationale, une collection de synthétiseur Text-To-Speech libres dans toutes les langues et dialectes possibles, pour Windows, Linux, Unix et Macintosh.

1. INTRODUCTION

Pourquoi avons-nous besoin de EULER ?

Les laboratoires de recherche publics et privés (universités et opérateurs télécom) ont investi des ressources considérables pour la mise au point de synthétiseurs de parole multilingues. Dans la plupart des cas, ces travaux de recherche non coordonnés ont généré des incompatibilités inter systèmes dues à un manque évident d'unification. Et ce malgré les outils et bases de données, connus et publiquement disponibles pour l'élaboration de système TTS. Chaque synthétiseur n'est qu'une implémentation de principes de base très similaires spécifiques à un laboratoire. De plus, la plupart des systèmes TTS multilingues ne sont en fait que des collections de TTS monolingues. Chaque TTS monolingue a été développé individuellement dans un laboratoire de la langue en question. Non seulement cette situation a un impact négatif sur les possibilités d'extensions d'un TTS vers d'autres langues, dialectes, accents, voix et styles de parole, mais en plus cela rend plus difficile l'intégration des TTS dans des produits finis (notamment pour les opérateurs télécom). En dernier lieu, le manque d'harmonisation dans la conception des TTS rend leur comparaison qualitative, module par module, très difficile à réaliser et cela restreint considérablement le déploiement des perfectionnements.

- A l'inverse de cette situation, des outils et des bases de données pour le développement de TTS multilingue ont été récemment, et de façon indépendante, mis à disposition par quelques universités européennes :La Faculté Polytechnique de Mons (FPMs) a contribué récemment au développement de synthétiseur multilingue « phonèmes vers parole » sous la forme du projet Internet MBROLA. L'objectif de ce projet est de favoriser les collaborations internationales quant à la réalisation de voix d synthèse dans nombre de langues et de dialectes. Ces voix sont gratuites, libres d'utilisation pour des applications non commerciales et non militaires. 19 langues existent et pour chaque langue, une ou plusieurs voix sont disponibles (26 voix en tout).
- L'Université d'Edimbourg (UED) a aussi largement contribué au développement de synthétiseurs parole à partir de texte, libre de droits pour applications non commerciales et non militaires au travers du projet FESTIVAL[FSSS]. FESTIVAL n'est rien de moins qu'une plate-forme de développement modulaire et générique, conçue dès l'origine dans des perspectives multilingues.
- L'université de Provence (UP) a été le coordinateur de la série de projets MULTEX [MUL] (LRE-MULTEXT, MULTEXT-EAST, MULTEXT-SW, MULTEXT-CATALOC, ALAF, etc.). Le but de ces projets est de développer des outils, des corpus et des ressources dans une large variété de langues. Ces outils et ressources sont disponibles gratuitement pour des applications non commerciales et non militaires.

L'intégration de ces contributions et de ces ressources dans une plate-forme commune est l'un des but fondamental de EULER.

2. UNE STRUCTURE OUVERTE ET MODULAIRE

2.1 Le Multi Layer Container

Dans les TTS, les données sont en général organisées en niveaux (orthographique, phonétique, prosodique, etc.). Pour que plusieurs modules, développés par des laboratoires différents, puissent communiquer ensemble, ils doivent au minimum partager les mêmes structures de données. Dans EULER, toutes les structures de données sont gérées par un objet partagé appelé «Multi Layer Container» (MLC) qui s'inspire des Multi-Level Data

Structures de Festival [FSSS] et de Speech Maker [SPMK].

Les avantages de la structure MLC par rapport à une description linéaire de données sont :

- Une lisibilité accrue des données et de règles. Les règles peuvent être par ailleurs implicites (c'est-à-dire spécifiées uniquement sur les données et les couches utiles).
- La traçabilité est facilitée : la recherche d'une donnée est simplifiée, grâce à un mécanisme de liens bidirectionnels entre données, et ceci aux travers des différentes couches.
- Les extensions sont aisées. La structure MLC accepte toujours de nouvelles couches de données. De nouveaux modules peuvent ajouter de nouvelles couches de données (accents toniques, balises de locuteurs pour les dialogues et/ou pour animation de visages, événements musicaux pour Karaoké, etc.). Ces nouvelles informations n'influençant pas les modules existants ; il est donc très facile d'étendre les fonctionnalités de EULER.

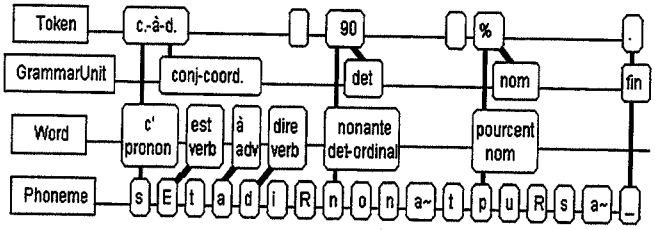


Figure1: phonétisation de la phrase «c.-à-d. 90 % »

La gestion de la MLC est intimement liée au noyau d'EULER.

2.2 La gestion des modules

Tous les modules disposent d'une interface commune. A l'initialisation, un fichier script définit la suite des modules nécessaires à l'architecture d'un TTS. Il est possible de définir plusieurs architectures en même temps : ceci permet par exemple de comparer immédiatement différents algorithmes. Bien entendu, certains modules sont communs aux différents TTS. Dans ce cas, ces modules sont partagés et ne sont chargés qu'une seule fois en mémoire. Grâce à cette gestion dynamique des modules, il est très facile de générer un ensemble de système de synthèse (correspondant à plusieurs voix, plusieurs styles de parole, plusieurs locuteurs, etc.) dans une seule et unique application.

2.3 Les modules

Chacun des modules est une unité autonome compilée. Néanmoins, ces modules peuvent être paramétrés et le noyau d'EULER leur transmet des arguments définis par l'utilisateur. La sémantique des arguments n'est pas figée. C'est à dire qu'un argument peut représenter une option de traitement (prononciation des nombres), le nom d'un lexique (dictionnaire d'exceptions) voire une fonction «callback» (animation de visage parlant) Une série d'arguments est transmise lors de la phase d'initialisation et une autre série est transmise à chaque appel du module. Ainsi, le comportement de l'application peut être modifié «au vol» (vitesse d'élocution par exemple).

Un des aspects intéressants lors la conception de modules est la généricité : si un algorithme est commun à plusieurs TTS (le module de synthèse, par exemple), il est plus pratique d'utiliser un module générique paramétré plutôt que de réaliser une série de modules spécifiques. Typiquement, le même module de synthèse peut produire une voix masculine ou féminine en fonction de ses paramètres.

2.4 Les moteurs

La sémantique des arguments des modules n'étant pas figé, elle peut aussi spécifier l'utilisation d'un moteur. Un moteur est une unité autonome compilée qui ne contient qu'un algorithme. Pour la phonétisation, par exemple, l'extraction des données graphémiques et grammaticales de la MLC est confiée au module, alors que la transcription graphèmes vers phonèmes est assurée par un moteur. Ce moteur peut être par exemple un algorithme basé sur des règles de réécritures régulières multi-niveaux (MLRR) ou une recherche dans arbre de décision (ID3). Le choix du moteur est réalisé dynamiquement par passage de paramètre.

3. LES MODULES EXISTANTS EN FRANÇAIS

3.1Pré-traitement

Cette étape couvre classiquement la recherche des fins de phrases, détection des abréviations, traitement des nombres, etc. Dans l'implémentation actuelle d'EULER, cette partie du traitement est confiée à un petit analyseur à base de grammaires régulières (détection des fins de phrases, décodage des nombres) et de lexiques (abréviations).

3.2 Lemmatisation

La tâche de lemmatisation consiste en une décomposition des mots en racine et suffixe. L'analyseur utilisé dans EULER s'appuie sur une base de données contenant plus de 32 600 lemmes et 160 formes de dérivation de la plupart des mots français. Cette base de données permet de retrouver toutes les natures grammaticales et les toutes les flexions d'environ 400 000 mots. Elle est dérivée du corpus français Vertex développé à Leuven par Piet Mertens. Le cœur de l'algorithme de lemmatisation est implémenté sous forme d'un moteur générique.

3.3 Analyse grammaticale

L'analyse grammaticale a été conçue sur une approche statistique. Le corpus initial comporte environ 4 300

phrases, soit 51 400 mots. Ce corpus a été initialement annoté par un analyseur grammatical automatique à base de règles. Ensuite, il a été complètement corrigé manuellement. En vue de réduire le nombre de classes grammaticales, les mots de ce corpus ont été regroupés en 39 classes (y compris les ponctuations). Finalement une grammaire a été extraite de ce corpus grâce aux outils de création de n-gram au format ARPA[ARP].

Pour la partie décodage, la phrase à analyser est représentée comme un treillis composé de la liste de toutes les classes grammaticales de chaque mot. Ensuite, un algorithme de type Viterbi donne le meilleur chemin par maximisation de la somme des vraissemblances associées aux séquences de *n* classes. Pour améliorer la partie décodage, un lexique de locutions prépositives et conjonctives est utilisé.

Sur les modules de lemmatisation et d'analyse grammaticale, les taux d'erreurs significatifs obtenus sur un échantillon de texte d'environ 2000 mots sont :

mot inconnu ou non reconnu	171 mots	8.5%
Déterminant pris pour nom	24	1%
Déterminant pris pour pro-clitique	74	3,6%
verbe pris pour nom	31	1,5%
adjectif pris pour nom	31	1,5%
adverbe pris pour nom	15	0,7%

Par la suite, on donne aux mots inconnus une «valeur» grammaticale équivalente aux noms. Ceci permet de diminuer fortement les erreurs de phonétisation.

3.4 Phonétisation

La tâche de phonétisation à partir des mots est assurée, pour le français, par l'algorithme ID3 qui réalise la prédiction par arbre de décision de la transcription phonétique à partir de vecteurs graphémiques et d'étiquettes grammaticales ceci afin de traiter les homographes hétérophones [Pagel][Black]. Le corpus d'entraînement utilisé contient environ 200 000 formes associées à leur analyse morphologique. Sur un ensemble de 1000 mots hors-lexique extraits du journal Le Monde, le système transcrit 91% de ces mots conformément à la référence (ce taux monte à 99,02 % sur les mots du lexique). Lors d'une évaluation auditive il s'avère que 95% des transcriptions sont acceptables.

Parallèlement à l'entraînement, cette méthode permet la compression de lexiques grâce à ses propriétés de généralisation ; sur ce corpus, il permet d'obtenir un rapport de compression de 22 pour 1.

D'autre part, un algorithme basé sur des règles de réécritures régulières multi-niveaux (MLRR) peut être utilisé pour les langues où l'on dispose de bases de règles phonétiques.

Ces algorithmes sont implémentés sous forme de moteurs; ils sont donc parfaitement interchangeables au sein du module sans modifier la structure du TTS.

3.5 Post-phonétisation

Ce petit module résout les liaisons françaises et l'élision des *e* muets. Il est basé sur un jeu de règles prenant en compte les caractéristiques graphémiques, grammaticales et phonétiques des données.

3.6 Prosodie

Le module de prosodie utilisé pour la génération du français est issu des travaux de F. Malfère [Malf]. Le système de génération de la prosodie est basé sur l'utilisation d'un corpus de parole analysé automatiquement. La segmentation phonétique et syllabique et l'évolution de la fréquence fondamentale sont extraites par le système MBROLIGN[MBR2], les caractéristiques grammaticales sont issues du système TTS LIPPS[LIPPS]. Le système de génération de la prosodie crée en premier lieu une représentation symbolique interne à partir de la syntaxe : les mots, leur nature grammaticale et leur transcription phonétique. Cette représentation intermédiaire sert de base à la génération du rythme et de la courbe de fréquence fondamentale. Les durées sont obtenues à l'aide d'un arbre de décision, entraîné sur les données alignées. La production de l'intonation est quant à elle basée sur un principe de concaténation de patrons intonatifs naturels, extraits du corpus d'entraînement, et correspondant aux groupes accentuels du français. Les patrons sont choisis automatiquement, de façon à minimiser une fonction de coût qui fait intervenir à la fois un coût de cible (ressemblance entre les caractéristiques des patrons disponibles et le patron dont on cherche à établir l'intonation) et un coût de concaténation (de façon à assurer une certaine continuité des courbes intonatives). Finalement, les données de durée et de fréquence fondamentale ainsi produites sont enregistrées dans la MLC.

3.6 Synthèse vocale

La génération des signaux acoustiques est confiée au synthétiseur MBROLA [MBR1], basé sur la concaténation de diphones. Ce synthétiseur est le cœur du projet MBROLA visant à l'élaboration d'une bibliothèque de voix par une politique de collaboration internationale.

3.7 Karaoké

L'importance de la généricité d'EULER a été clairement mise en évidence pour la création d'une application de Karaoké de synthèse. Des modules adaptés à la lecture de fichiers Karaoké (.kar) ont été développés assurant les rôles d'extraction de données de paroles et de mélodie de fichiers Karaoké et d'enregistrement de celles-ci dans

la MLC. Puis, à partir de ces données, des modules existants sont directement utilisés (phonétisation, synthèse) en vue de produire de la parole chantée. Cette application, basée sur EULER, a été développée en moins d'un mois.

4. Conclusion et perspectives

La vocation du projet EULER est de proposer cette plate-forme de recherche et de développement à toute la communauté scientifique. Euler est gratuit et libre d'utilisation dans le cadre de la recherche (utilisation non commerciale et non militaire). Une partie des sources est même publique sous licence GNU (le noyau et la plupart des modules), ceci afin d'éviter une situation où un développeur de modules ou d'applications se retrouverait irrémédiablement soumis au bon vouloir des initiateurs du projet. Les codes sources, écrits en C++ ont été développés, pour les plates-formes Windows, Unix/Linux et seront bientôt portés sur Macintosh.

Tout comme le projet MBROLA, le but du projet EULER est de réunir une collection de TTS dans le plus grand nombre de langues et dialectes possibles, libre d'utilisation non commerciale et non militaire. Un appel à collaboration internationale sera prochainement lancé pour atteindre ce but.

La protection de la propriété intellectuelle des développeurs est assurée sur plusieurs niveaux : les moteurs, modules et bases de données, sont systématiquement fournis avec des marques de copyright et licences appropriées. Les applications logicielles utilisant EULER affichent automatiquement ces informations au lancement de l'application.

Actuellement, deux langues sont disponibles : le français l'arabe ; l'anglais est en cours de finalisation. La ré-utilisabilité des composants EULER permet de générer rapidement, à partir de la première version française, les langues arabe et anglaise.

5 Remerciements

Les auteurs tiennent à remercier tous ceux qui ont contribuer au développement de ce projet : Piet Mertens pour ses outils Morlex et Vertex, Richard Beaufort pour sa contribution au dictionnaire phonétique français, BABEL Technologies S.A. pour son engagement en faveur de la mise à disposition d'outils gratuites à des fins non commerciale, Alan Black pour sons soutient psychologique lors de l'élaboration de la MLC, et les gens de l'Université d'Aix-en-Provence pour leurs encouragements lors de la genèse de EULER .

Bibliographie

[EUL] The EULER Project : *http://tcts.fpms.ac.be/synthesis/euler/*

[MBR1] T.Dutoit, V. Pagel, N. Pierret, F. Bataille, O. Van Der Vrecken, "The MBROLA project : towards a set of high quality speech synthetizers free of use for non commercial purpose", icslp'96 *http://tcts.fpms.ac.be/synthesis/*

[FSSS] A.W. Black, P. Taylor and R. Caley, "The Festival Speech Synthesis System", University of Edinburgh, 1997 *http://www.cstr.ed.ac.uk/projects/festival*

[SPMK] Van Leeuwen & E. Te Lindert, "Speech Maker: a flexible and general framework for text-to-speech synthesis, and its application to Dutch", Computer Speech and Language, pp. 149-167, 1993.

[MUL] J. Veronis, D. Hirst, R. Espesser and N. Ide, "NL and speech in the MULTEXT project", AAAI'94 Workshop on Integration of Natural Language and Speech, 1994.

[ARP] Ronald Rosenfeld, Philip Clarkson, "SLMT ToolKit", Carnegie Mellon University, Cambridge university, *http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html*

[Malf] F. Malfrère, T. Dutoit, P. Mertens, 1998, "Un générateur de Parole "Tout Automatique"" Proc. XXIIèmes Journées d'Etudes sur la Parole, Martigny, pp. 147-150.

[Pagel] V. Pagel, K. Lenzo, A. Black, 1998, "Letter-to-Sound Rules for Accented Lexicon Compression", Proc. ICSLP'98, Sydney, Australia, pp. 252-255.

[Black] A. Black, K. Lenzo, V. Pagel, 1998, "Issues in Building General Letter to Sound Rules", Proc. 3rd ESCA/COCSADA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 77-81.

[MBR2] F.Malfrère and T.Dutoit, "High Quality Speech Synthesis for Phonetic Speech Segmentation", Proceedings of EuroSpeech'97, pp. 2631-2634, 1997.

[LIPSS] T.Dutoit, "High Quality Text-to-Speech Synthesis for the French Language", Ph. D. dissertation, Faculté Polytechnique de Mons, 1993.

Amélioration automatique de l'intelligibilité de la parole

Vincent Colotte et Yves Laprie

LORIA
BP 239 - Campus scientifique - 54506 Vandœuvre-lès-Nancy, FRANCE
Mél: Vincent.Colotte@loria.fr , Yves.Laprie@loria.fr

RÉSUMÉ

This paper presents a speech signal transformation which slows down speech signals selectively and enhances some important acoustic cues. This transformation can be used not only for hearing aids but also for second language acquisition by facilitating oral comprehension.
The strategy used to control slowing down exploits a spectral variation function which locates rapid spectral changes. The enhancement simply consists of amplifying stop bursts and unvoiced fricatives. These acoustic cues are detected automatically through the examination of energy criteria.
This approach was evaluated in the context of second language acquisition. Experiments show that the oral comprehension is improved.

1. INTRODUCTION

Le papier présente une méthode de transformation de la parole permettant de ralentir sélectivement son débit et de renforcer les évènements transitoires, pour améliorer l'intelligibilité. En effet, le ralentissement peut améliorer la compréhension orale lors de l'apprentissage de langues étrangères ou compenser la perte de sélectivité temporelle chez les personnes âgées notamment. Le ralentissement peut être appliqué globalement ou sélectivement sur certaines parties du signal (i.e. sur certains indices acoustiques). La dernière solution a l'avantage de ne pas allonger exagérément la durée du signal. De plus, il est possible de combiner le ralentissement par un renforcement des évènements transitoires dans le but d'améliorer la perception de certains indices acoustiques. De même que le ralentissement et le renforcement favorisent la discrimination des sons entendus, ces techniques rendent le signal plus robuste aux dégradations possibles (bruit, canal de télécommunications...).
Différentes méthodes en synthèse de la parole peuvent être mises en œuvre pour effectuer le ralentissement et le renforcement. Nous travaillons dans le contexte de la modification d'un signal de parole avec TD-PSOLA (*Time Domain Pitch Synchronous Overlap-Add*) [MC90] qui est une méthode bien connue en synthèse de la parole pour son faible coût calculatoire. PSOLA repose sur une décomposition du signal temporel en fenêtres recouvrantes synchronisées sur la fréquence fondamentale. Si cette méthode nécessite peu de calculs, elle nécessite en revanche la connaissance des marques de pitch indiquant le centre des fenêtres. L'algorithme de marquage du pitch que nous proposons exploite les résultats de l'algorithme d'extraction de la fréquence fondamentale. Nous effectuons la propagation des marques de période en période à l'aide d'un algorithme de programmation dynamique; ainsi le marquage est optimal sur l'ensemble du signal. De plus, cette méthode a l'avantage d'être automatique et indépendante de l'algorithme d'extraction de la fréquence fondamentale (pour plus de détails, voir [LC98]).
La stratégie de contrôle du débit dépend du niveau auquel on veut modifier la parole: au niveau prosodique le ralentissement permet de renforcer l'intelligibilité d'un groupe de mots ou d'une partie de la phrase alors qu'au niveau phonétique le ralentissement porte sur des sons précis pour en améliorer l'intelligibilité, et ainsi améliorer l'intelligibilité de la phrase entière. A. Nakamura et al. [NSI+96] ont travaillé de manière à permettre aux personnes âgées de mieux comprendre les bulletins d'information télévisés. L'amélioration de l'intelligibilité est basée sur la modification des "groupes de souffle", et intervient donc au niveau prosodique. Il a été montré, pour les bulletins d'information en japonais, que la mélodie (en l'occurrence la fréquence fondamentale) suit une ligne de déclinaison sur les groupes de souffle et que l'information principale se situe surtout dans les parties à pitch élevé, c'est-à-dire au début des groupes. C'est pourquoi le signal est ralenti pendant la première partie des groupes de souffle puis accéléré sur la fin. Les silences entre groupes de souffle sont eux aussi accélérés. Dans les deux cas, l'accélération est motivée par le souci de ne pas trop allonger la durée du signal, l'opération s'effectuant en temps réel. V. Hazan et A. Simpson [HS98], quant à eux, travaillent au niveau phonétique; ils recherchent, sur le signal, les régions à forte densité d'indices acoustiques puis renforcent ces régions, principalement par filtrage, pour améliorer l'intelligibilité. Ces régions indiquent les transitions du signal, c'est-à-dire les endroits où les caractéristiques acoustiques du signal changent très rapidement. Ces régions correspondent aux *landmarks* introduites par S.A. Liu [Liu96] pour repérer les régions de transition (ouverture/fermeture de la glotte, début/fin d'un burst, d'une nasale...). S.A. Liu assimile les changements de caractéristiques acoustiques aux variations de l'énergie du signal sur plusieurs bandes de fréquence: à partir de connaissances expertes, elle associe un type de variation acoustique à un type d'évènement articulatoire.
Dans la première partie de ce papier, nous présentons

la mise en œuvre et la stratégie de contrôle du débit à partir du calcul d'une fonction d'évaluation des variations spectrales. Dans une deuxième partie, nous décrivons la stratégie de renforcement qui repose sur la détection de bursts et de fricatives sourdes. Nous conclurons sur les résultats obtenus et les perspectives envisagées.

2. RALENTISSEMENT

2.1. *Mise en œuvre du marquage du pitch et du ralentissement*

TD-PSOLA est une technique qui permet de modifier facilement le débit de la parole et le contour de la F0 ; son principal avantage est son faible coût calculatoire. Contrairement à la synthèse à partir du texte où le problème est de concaténer de courtes unités de parole, notre objectif est de modifier la totalité de la phrase. TD-PSOLA est facile à mettre en œuvre du moment que la localisation des marques du pitch, qui décomposent le signal en fenêtres recouvrantes synchronisées sur la fréquence fondamentale, est réalisée. Comme il existe un grand nombre d'algorithmes de détection de la fréquence fondamentale qui deviennent de plus en plus robustes, il semble intéressant d'exploiter les résultats de l'extraction du pitch pour déterminer les marques du pitch. Une fois les marques du pitch connues, la mise en œuvre du ralentissement devient simple.

Déterminer les marques du pitch à partir de la connaissance des résultats de la fréquence fondamentale consiste à sélectionner des extrema du signal séparés d'un intervalle correspondant à la période du pitch locale. Les marques sont choisies parmi les extrema locaux (minima ou maxima suivant le meilleur coût global). La programmation dynamique permet de propager les marques sur l'ensemble de la phrase de manière optimale. Le marquage du pitch s'effectue en deux étapes.
La première étape consiste à construire un ensemble de candidats à partir des extrema locaux. Les candidats sont recherchés sur des fenêtres espacées régulièrement. Pour être assuré que toutes les périodes de pitch ont un candidat au moins, la fenêtre de recherche doit être plus petite que la plus petite période de pitch sur toute la phrase à modifier.
La deuxième étape consiste à choisir, parmi les candidats, un ensemble, d'extrema séparés par une période de pitch. Soit $C = [c(i)] = c(1) \dots c(i) \dots c(N)$ l'ensemble des candidats où $c(i)$ est l'instant du i^{eme} extremum. Nous devons déterminer une fonction de sélection j donnée par $J = [j(k)] = j(1) \dots j(k) \dots j(K)$ avec $K < N$ et telle que $j(k) < j(k+1)$ (pour préserver l'ordre chronologique). Une solution pour le marquage du pitch est alors $\overline{C} = [c(j(k))] = c(j(1)) \dots c(j(k)) \dots c(j(K))$. Pour trouver la meilleure fonction de sélection nous devons définir un critère qui exprime la qualité des marques du pitch. Nous avons choisi le critère local suivant pour deux marques consécutives :

$$d(c(i),c(l)) = |(c(l) - c(i)) - pitchPeriod(c(i))| - \alpha \times |amplitude(c(i))| \quad (1)$$

Le premier terme exprime le fait que la distance entre les marques l et i est approximativement égale à la période de pitch ; le second terme favorise les forts extrema et α représente le compromis entre les deux termes. La recherche de l'ensemble des marques revient donc à trouver K_{opt} et j_{opt} qui minimisent $D = \sum_{k=1}^{K-1} d(c(j(k)),c(j(k+1)))$. Ce problème est résolu par programmation dynamique (voir [LC98]) et les résultats obtenus sont très bons. D'un point de vue pratique, le marquage du pitch est effectué pour les minima et les maxima séparément et la meilleure solution est conservée.
Modifier le débit de la parole revient à dupliquer les fenêtres centrées aux marques du pitch sans changer le pitch. Pour connaître les fenêtres à dupliquer nous utilisons des marques virtuelles qui représentent les positions de marques pour le taux du débit cible. Soit $c_a(i)$ la i^{eme} marque d'analyse (trouvée par l'algorithme précédemment exposé), et $c_v(i)$ la i^{eme} marque virtuelle, $c_v(i+1)$ est donnée par $c_v(i+1) = c_v(i) + r(i+1) \times (c_a(i+1) - c_a(i))$ où $r(i+1)$ est le taux de modification du débit appliqué entre l'instant $c_a(i)$ et $c_a(i+1)$.

2.2. *Stratégie de contrôle du ralentissement*

L'approche la plus pertinente pour une aide aux malentendants ou pour l'apprentissage des langues étrangères semble se situer au niveau phonétique contrairement à [NSI+96] qui opère au niveau prosodique. En effet la méthode proposée par A. Nakamura et al. [NSI+96] ne semble pas transposable au français. Bien que la ligne de déclinaison existe aussi en français, l'importance de l'information apportée par les groupes de souffle ne suit pas la déclinaison de la fréquence fondamentale comme cela semble être le cas dans les bulletins d'information télévisés japonais. De plus, les modifications au niveau phonétique permettent de focaliser le ralentissement sur des évènements particuliers : seulement un type de son, difficile à percevoir, peut être transformé. La stratégie de ralentissement doit donner lieu à un algorithme automatique et doit pouvoir être piloté par la détection de certains indices acoustiques.

Deux directions sont alors possibles. Soit le signal est segmenté selon les frontières phonétiques de la phrase énoncée - il s'agit donc d'une segmentation en sons -, soit le signal est marqué aux instants[1] à forte densité d'indices acoustiques, c'est-à-dire les régions où les caractéristiques acoustiques du signal varient fortement et rapidement.
Une segmentation phonétique stricte peut être obtenue implicitement lors de la reconnaissance automatique de la parole ou à partir d'une transcription phonétique manuelle : ces deux solutions ont été écartées, la première pour sa lourdeur en calcul et sa précision, et la dernière, bien qu'envisageable dans des exercices de compréhension orale lors d'apprentissage de langue, car nous voulons mettre en œuvre une méthode totalement automatique.
La seconde approche repose sur la localisation des ré-

1. Une région, ici, est définie par un intervalle centré sur un instant. La région est localisée par cet instant et les modifications sont effectuées sur son voisinage.

gions à forte densité d'indices acoustiques. Contrairement à Liu [Liu96] qui utilise un critère sur l'énergie et une connaissance experte sur les caractéristiques des évènements à localisés, nous utilisons une méthode qui évalue les variations acoustiques de la parole. Cette méthode, appelée *Spectral Variation Function*, proposée G. Flammia et al. [FDAL92] et F. Brugnara et al. [BMGO92] utilise une analyse mel-cepstre. Un coefficient, reflétant le taux de variation du spectre, est calculé pour chaque fenêtre (de 20 ms toutes les 10 ms) par rapport aux fenêtres voisines. Le recherche des maxima locaux de la fonction SVF nous donne les instants de forte variation des caractéristiques acoustiques. Nous avons retenu cette méthode qui indique les régions à forte variation acoustique parce qu'elle permet de localiser 82% des frontières de sons placées par un expert. Les 18% restant sont soit des marques mal placées (à plus de 20 ms) soit des insertions. Ce deuxième type d'erreurs n'est pas trop gênant voire utile ; en effet, étant donné le but à atteindre, à savoir ralentir le signal pour une meilleur intelligibilité, le fait de prendre plus de marques et de ralentir la parole au voisinage de ces marques ne risque pas de diminuer l'intelligibilité, d'autant que les marques sont proches et donnent lieu de fait à un seul ralentissement sur une région englobant ces marques. Le choix de la valeur du facteur de ralentissement est arbitraire, mais une valeur trop forte (supérieure à 3) dénature le son par rapport à l'articulation habituelle (en particulier, les bursts peuvent être artificiellement transformés en fricatives). Nous avons donc retenu une valeur de 1.8 à 2 pour ce facteur. Même avec cette valeur de ralentissement plutôt élevée, l'allongement moyen global (sur toute la phrase) est seulement de 1.3.

3. RENFORCEMENT

La recherche des segments à renforcer s'est principalement focalisée sur les occlusives et les fricatives. D'après V. Hazan et A. Simpson [HS98], le renforcement de ce type de phonème permet d'améliorer l'intelligibilité en parole spontanée. Nous avons décidé de nous focaliser sur les bursts et fricatives sourdes car ces sons peuvent être localisés avec un fort degré de robustesse et leur renforcement améliore la perception de la structure temporelle de la parole.
Nous allons expliquer comment les fricatives et les bursts sont détectés à partir d'un critère basé sur l'énergie. Différencier une fricative d'un autre son peut être facilement réalisé à partir de l'énergie. En première approximation, l'énergie d'une fricative est principalement localisée en haute fréquence. Nous avons donc choisi de calculer le rapport de l'énergie dans la bande 3600 – 6000 Hz sur l'énergie dans la bande 600 – 1000 Hz, les autres fréquences étant considérées comme moins pertinentes. Les fricatives sourdes correspondent à une valeur élevée de ce rapport.
Les occlusives se caractérisent par l'absence d'énergie pendant l'occlusion : cela se répercute par une faible moyenne d'énergie au centre du segment de l'occlusion par rapport à l'énergie aux bords. En présence d'une occlusion, la moyenne au centre est plus faible que la moyenne globale. Le second indice, qui différencie un burst d'un début de voisement, par exemple, est la dérivée de l'énergie. Un burst présente un pic dû à la forte variation du spectre au moment de l'explosion. Un seuil est utilisé pour sélectionner les pics significatifs (50% du maximum de l'amplitude de la dérivée). Fig. 1 résume la stratégie utilisée pour détecter les bursts et les fricatives.

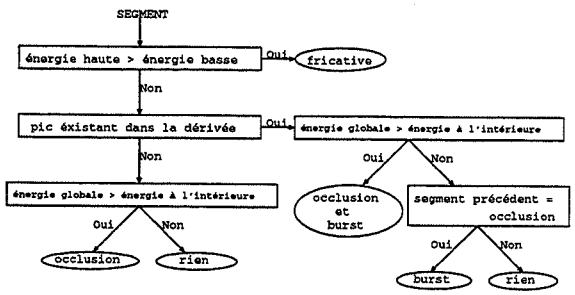


FIG. 1 – *Algorithme simplifié de détection des occlusives et des fricatives*

La stratégie de renforcement de l'énergie du signal est basée sur les expériences de V. Hazan et A. Simpson [HS98]. Le principe est d'amplifier progressivement les transitions des bursts et des fricatives (dans le domaine temporel) jusqu'à un niveau voulu et de revenir au niveau initial (voir Fig. 2).

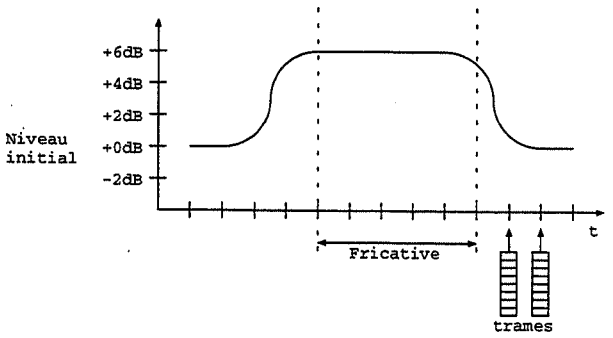


FIG. 2 – *Exemple de renforcement d'une fricative sourde*

4. EXPÉRIMENTATION

Comme mentionnées dans l'introduction, les deux applications possibles sont l'aide aux malentendants et la compréhension orale pour une seconde langue. Tout d'abord, nous avons testé l'amélioration de la compréhension orale (de phrases anglaises par des personnes françaises) car l'ajustement de la stratégie de contrôle du débit pouvait se faire facilement dans ce cas.
Considérant que nos transformations s'apparentent à des études de perception basées sur l'exagération d'indices acoustiques, nous avons décidé d'évaluer l'amélioration au niveau des mots plutôt que sur des unités plus spécifiques (de type VCV par exemple).
Le corpus de test est constitué de 50 phrases sélectionnées dans la base de données TIMIT.

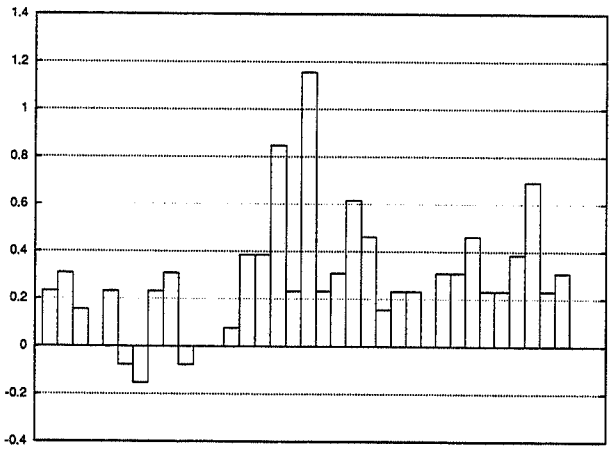
4.1. Evaluation des transformations

La première évaluation porte sur la pertinence du ralentissement sélectif et du renforcement. Chaque technique donne séparément de bons résultats et sont robustes aux erreurs. En effet, les erreurs commises par le marquage SVF sont principalement des insertions et aucun compromis n'est donc fait par rapport à l'in-

telligibilité. De même, les erreurs commises lors de la détection des bursts et des fricatives sont principalement des omissions de faibles bursts ou fricatives, ce qui là encore ne détériore pas l'intelligibilité de la parole, contrairement à des insertions (dues à de fausses détections) qui auraient pu introduire des bursts ou des fricatives artificiels.
Les 50 phrases du corpus de test ont été transformées et évaluées. Nous n'avons trouvé que deux bursts artificiels, dont un masqué par une fricative voisine (ce qui ne change pas la perception du mot) et l'autre perçu comme un clic. Nous avons remarqué une erreur de marquage de pitch sur un début de voisement d'un burst non-voisé [d] qui décale la portée de l'amplification et modifie la perception. Comme attendu, les marques SVF apparaissent dans des régions qui contiennent des variations spectrales rapides (principalement aux transitions formantiques et en bord de nasales). En général, plusieurs marques sont détectées dans les régions correspondantes à de rapides transitions formantiques. Comme ces marques sont très proches les unes des autres, elles donnent lieu à un taux de ralentissement unique et ne perturbent pas la stratégie de contrôle du débit.

4.2. Evaluation perceptive

13 adultes français ont participé à deux sessions d'expérimentation d'une demi-heure. Dans la première, les 50 phrases sont les phrases originales. Dans la seconde session, 25 furent conservées et 25 autres furent modifiées par la transformation exposée précédemment. Le corpus de test a été aléatoirement mélangé et les personnes devaient compléter le ou les deux mots manquants dans la transcription des phrases qu'ils écoutaient. Nous avons considéré quatre niveaux de réponse : **0** aucune réponse n'a été donnée, **1** la réponse donnée n'a rien en commun avec le mot correct, **2** au moins la moitié de phonèmes sont corrects, **3** le mot a été bien reconnu. Tab. 1 donne la différence



TAB. 1 – *Différences des valeurs d'identifications pour les 37 mots cibles*

en moyenne (sur les 13 personnes) pour les 37 mots testés entre leur version modifiée et leur version originale. Un test de Student a montré que les valeurs d'identifications ont été améliorées significativement ($p < 0.02$). Les résultats ont montré par ailleurs que l'amélioration est uniformément distribuée entre les bursts et les fricatives et que le ralentissement des transitions ne change pas la perception des transitions, excepté pour un seul burst dont l'articulation a été modifié et qui a été ainsi perçu comme une fricative.

5. CONCLUSION

La principale force de notre approche est le fait qu'elle repose sur une stratégie simple : renforcement des occlusives et des fricatives sourdes fiables. De plus, les modifications portent sur les régions ou les phonèmes qui ont un effet significatif sur la compréhension : les transitions - régions contenant une forte concentration d'indices acoustiques - et les phonèmes de type occlusives et fricatives. La combinaison du ralentissement sélectif à partir des marques SVF et du renforcement acoustique des bursts et des fricatives améliore l'intelligibilité de la parole. Les résultats se trouvent sur *http://www.loria.fr/~colotte* [2].
L'avantage de notre approche est qu'elle est totalement automatique ; ainsi, elle peut être facilement combinée avec un système de synthèse de parole visuel dans le but de compléter le signal acoustique avec les mouvements des lèvres pour exploiter l'effet McGurk [MM76].

RÉFÉRENCES

[BMGO92] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo. Improved connected digit recognition using spectral variation functions. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 627–630, Banff, Canada, 1992.

[FDAL92] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg. Segment based variable frame rate speech analysis and recognition using a spectral variation function. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 983–986, Banff, Canada, 1992.

[HS98] V. Hazan and A. Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226, 1998.

[LC98] Y. Laprie and V. Colotte. Automatic pitch marking for speech transformations via td-psola. In *IX European Signal Processing Conference*, Rhodes, Greece, 1998.

[Liu96] S.A. Liu. Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.*, 100(5):3417–3430, November 1996.

[MC90] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, 1990.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 246:745–746, 1976.

[NSI+96] A. Nakamura, N. Seiyama, A. Imai, T. Takagi, and E. Miyasaka. A new approach to compensate degeneration of speech intelligibility for elderly listeners. *IEEE Trans. on Broadcasting*, 42(3):285–293, September 1996.

2. Ils sont susceptibles d'être modifiés en fonction de nos tests perceptifs.

Évaluation des systèmes d'analyse-modification-synthèse de parole

Gérard Bailly

Institut de la Communication Parlée - UMR CNRS $n°5009$ - INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél. : ++33 04 76 57 47 11 - Fax : ++33 04 76 57 47 10
e-mail : bailly@icp.inpg.fr - http ://www.icp.inpg.fr/

Abstract

The evaluation of analysis-modification-synthesis systems provided in the literature is often informal and results cannot be compared. We propose here a methodology for comparing the performance of such systems using common ressources. The proposed benchmark consists in performing various prosodic transplantations from a given natural utterance to various prosodic "variants" of the same sentence. These prosodic "variants" have been recorded by the same speaker in order to obtain both credible prosodic transplantations and reference target signals : we evaluate the systems by estimating their capacity in reproducing the entire spectro-temporal structure of the natural targets given only the source signals and the prosodic modifications to be performed. These principles and first results obtained within the framework of the Cost258 project are presented.

Introduction

La plupart des systèmes de synthèse de parole utilise la concaténation d'unités stockées pour générer un continuum sonore dont certaines caractéristiques seront alors modifiées en fonction de consignes prosodiques (voir figure 1). Le travail est donc partagé entre l'algorithme de sélection des unités et le système d'analyse-modification-synthèse de signal (appelé "codeur" dans la suite). Bien qu'un algorithme de sélection opérant sur de larges dictionnaires d'unités peut réduire la distorsion à effectuer par le codeur en incorporant dans son processus de sélection la minimisation de la distance des caractéristiques prosodiques des signaux sélectionnés à la cible désirée [5, 14], il est cependant difficile d'obtenir une qualité de voix homogène sur de larges corpus et il semble difficile d'éviter le recours à un codeur ne serait-ce que pour effectuer un mise à l'échelle ou un lissage de certaines caractéristiques spectrales.

La phase de modification des caractéristiques prosodiques est cruciale : la modification d'une caractéristique aussi standard que la mélodie est accompagnée dans les signaux naturels d'une covariation de l'ensemble de la représentation paramétrique du signal. Ainsi les variations de la forme de l'onde glottique ou des formants en fonction de la fréquence fondamentale ou de l'effort vocal ont été bien étudiés [10] et utilisés par les systèmes de synthèse par règles. Dans le cas de codeurs sans accès explicite à ces paramètres fins, les covariations doivent être implicites. Parmi ces propriétés implicites, on trouve par exemple l'invariance de forme, propriété commune à de nombreux codeurs (TDPSOLA [6], codeurs sinusoïdaux [1, 15] ...) qui préserve la forme globale de l'onde malgré les changements de fréquence fondamentale. La préservation de ces propriétés implicites n'est cependant ni nécessaire ni suffisante pour garantir la cohérence des modifications et il semble difficile pour l'instant de s'affranchir d'une évaluation comparative des codeurs.

Cet article propose une méthodologie d'évaluation des codeurs de synthèse basée sur des tâches de transplantation prosodique de complexité croissante.

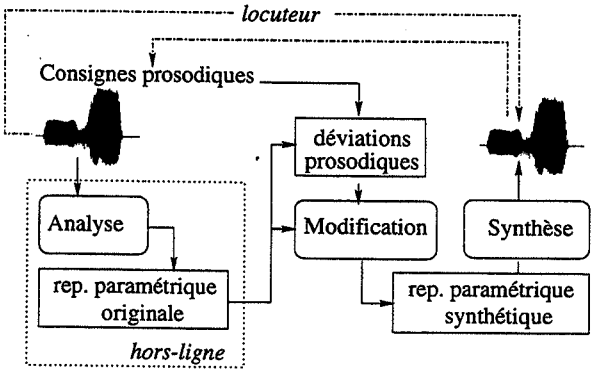


Fig. 1: Schéma de principe d'un codeur. Grâce à une méthode d'analyse spécifique, on délivre une représentation de signaux originaux qui sera modifiée pour obéir aux consignes prosodiques calculées par le processus de synthèse. Le principe d'évaluation proposé est figuré en traits discontinus : un même locuteur prononce les signaux sources et cibles. Le codeur est alors soumis à une tache de transplantation prosodique.

1. Évaluation des codeurs : un problème mal-posé

Du fait de l'émergence des méthodes statistiques en synthèse de parole, de la (relative) facilité de mise en oeuvre de systèmes de synthèse par concaténation, un système de synthèse n'est plus nécessairement le fruit d'un travail acharné de quelques experts phonéticiens. Grâce à des outils génériques [9, 19] et des ressources de qualité disponibles, il est facile de construire des systèmes opérationnels. On comprend dès lors la nécessité d'élaborer des méthodologies d'évaluation modulaire ("glass box") afin d'identifier les propriétés intrinsèques des briques du lego communautaire. Et la

littérature ne nous fournit que peu d'expériences de ce type : la plupart des évaluations impliquent des systèmes complets (d'ailleurs souvent identifiés par le codeur utilisé [16]) ou comparent les codeurs au sein d'une même architecture [8, 17, 18]. Outre le fait que ces comparaisons n'incluent pas une échelle de référence (par exemple, de la parole naturelle ou au moins de la prosodie naturelle), il est impossible d'établir une grille de lecture des résultats tant les propriétés des codeurs sont différentes en regard des tâches qui leur sont demandées : ainsi TDPSOLA est très sensible aux problèmes de concaténation (phases, structure spectrales aux points de concaténation) mais permet de préserver avec grande précision des zones du signal difficiles à analyser avec des codeurs paramétriques [4]. D'autre part, les propositions de codeurs ne sont accompagnées le plus souvent que d'évaluation informelle impliquant les codeurs dans des tâches peu réalistes (manipulation de durées ou de fréquence fondamentale uniforme de signaux naturels) sur des stimuli ad hoc [1]. L'auditeur est ainsi biaisé vers un jugement esthétique sans référence aux performances attendues.

Les évaluations subjectives sont lourdes - et donc précieuses - et présentent une photographie d'un système à un instant donné de son développement. Il y a donc nécessité de capitaliser les résultats obtenus de manière à permettre aux chercheurs d'utiliser l'évaluation comme outil de diagnostic et d'enrichir l'éventail des tâches de modification et des méthodologies d'évaluation afin de mettre en valeur les propriétés attendues des codeurs : un système de synthèse par concaténation de larges unités ayant peu recours à la modification insistera sur la transparence du processus d'analyse-synthèse alors que des systèmes de synthèse multi-style faisant largement appel à des modèles de variabilité mettra l'accent sur la versatilité du codeur.

2. Identification de tâches

Alors que la recherche en reconnaissance de parole, identification de la langue et du locuteur est largement structurée et dynamisée par des campagnes d'évaluation internationales, l'évaluation des systèmes de synthèse reste un sujet controversé. Peu de campagnes d'évaluation permettent d'effectuer un réel diagnostic des principales lacunes des modules utilisés dans les synthétiseurs actuels ou de leur mode de couplage. Notons cependant l'efficacité d'une telle campagne lorsque les bonnes grilles de lecture des résultats sont identifiées [21].

2.1. *Tâches élémentaires*

La grille de lecture des performances des codeurs proposée s'appuie sur l'identification de tâches élémentaires auxquelles les codeurs sont soumis et sur lesquelles une différenciation nette est attendue. Une liste non exhaustive comprendra d'abord la manipulation des paramètres prosodiques classiques : la fréquence fondamentale, la durée et l'intensité des sons, puis la manipulation du timbre pour diverses applications telles que lissage aux points de concaténation, génération multi-style ou transformation de voix. Il convient de plus d'évaluer des manipulations sur divers types de sons : ainsi, la manipulation du fondamental de fricatives voisées pose le problème de la synchronisation entre bruit de friction et oscillation glottique [13] ; de même la manipulation des durées de sons non-voisés pose le problème de la définition et du traitement des trames à beaucoup de codeurs synchrones avec le fondamental.

2.2. *Vers des tâches complexes*

Comme on l'a vu dans l'introduction, ces manipulations ne sont pas totalement contrôlées de manière indépendante et doivent être effectuées de manière synergétique pour structurer le discours. Une évaluation doit donc comprendre une complexification progressive des tâches de manière à tester non plus le module d'analyse-synthèse mais aussi le module de modification des représentations qui prend en charge la négociation des consignes.

2.3. *Ressources*

Signaux La comparaison des codeurs à une référence absolue naturelle suppose de disposer de ressources source et cible illustrant ces manipulations élémentaires et complexes. Le locuteur doit ainsi être instruit des tâches soit *de manière implicite* par usage de stimuli synthétiques guidant ses réalisations soit directement par imitation soit indirectement en donnant des buts à atteindre (Barbosa [3] a par exemple suggéré des variations de débit par des questions synthétiques à la place du métronome souvent utilisé) ; soit *de manière explicite* par description textuelle d'une situation propre à suggérer une intonation voulue.

Descripteurs Il convient de disposer de descripteurs prosodiques de référence afin de pouvoir comparer les codeurs dans les même tâches de transplantation prosodique. Ces descripteurs doivent donc au minimum comprendre : un marquage de périodes, un marquage des frontières phonémiques. Ils peuvent et doivent ensuite être enrichis par des descripteurs "intelligibles" i.e. sur lesquels une connaissance explicite puisse être acquise (pente spectrale, coefficients de description de l'onde glottique, jitter, shimmer ...).

3. L'exemple du serveur Cost258

Dans le cadre du Cost258 [2], une première version d'un serveur de ressources [3] pour l'évaluation des codeurs a été mis en place.

3.1. *Les ressources*

Plusieurs locuteurs ont rempli les tâches suivantes : (1) VO (*hauteurs vocaliques*) : les 10 voyelles orales du Français ont été prononcées à plusieurs hauteurs de

[1] rendons grâce à Speech Communication qui permet de proposer en ligne les stimuli utilisés par les auteurs (voir par exemple [20]) dans `www.elsevier.nl:80/inca/publications/store/5/0/5/5/9/7`.

[2] `www.unil.ch/imm/docs/LAIP/COST_258/cost258.htm`

[3] `www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html`

manière à couvrir une plage d'environ 3/4 d'octave; (2) FD (*durées*) : deux versions courtes et longues des 6 fricatives du Français ont été produites isolément. (3) AT (*attitudes*) : 6 phrases ont été prononcées avec 6 attitudes prosodiques différentes. (4) EM (*émotion*) : 1 phrase a été prononcée avec 4 émotions différentes.

Chaque locuteur a aussi prononcé chaque énoncé avec une intonation et un rythme les plus plats possible, de manière à fournir un stimuli source le plus proche possible de la base fournie par un système par concaténation idéal (voir par exemple la normalisation opérée par le système MBROLA). Tous les signaux ont été échantillonnés à 16kHz, segmentés et les marqueurs de périodes positionnés semi-automatiquement. Le centre de réalisation des phonèmes a été marqué et l'énergie à court-terme de ce centre a été ajouté au jeu de descripteurs prosodiques.

3.2. Les codeurs de référence

Plusieurs codeurs ont accompli les diverses tâches de transplantation : nous avons implémenté une version éprouvée de TDPSOLA [2]. 4 autres codeurs fournis par d'autres laboratoires partenaires ont traité ces stimuli (c1_0, c2_0, c3_0, c4_0). Trois d'entre eux ont de plus fourni une nouvelle version (c1_1, c2_1, c4_1) suite à la première campagne effectuée. Trois autres codeurs ont été simulés en additionnat à la cible divers niveaux de bruit additionnel (30dB, 20dB, 10dB).

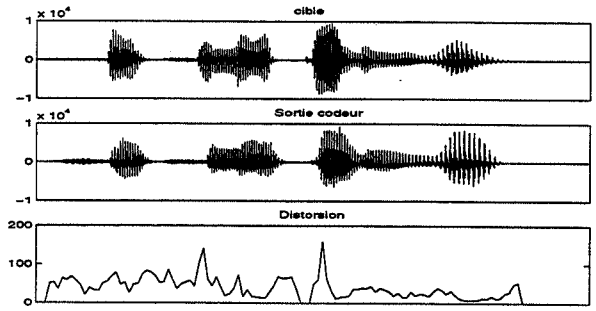


FIG. 2: Évolution de l'erreur de modélisation par la mesure de distorsion WSS.

4. PROCÉDURES D'ÉVALUATION

Un tel serveur vise à fournir des ressources de référence aux développeurs de codeurs. Il vise aussi à fournir une base méthodologique d'évaluation des systèmes.

4.1. Évaluation objective

De nombreux travaux s'attachent à reproduire par des mesures de distorsion les scores moyens d'opinion (MOS) obtenus lors d'évaluations subjectives telles que l'estimation de la gêne perceptive occasionnée par des discontinuités spectrales aux points de concaténation entre deux segments de parole [7, 14] ou la prédiction plus globale des performances d'algorithmes de réhaussement de parole [11].

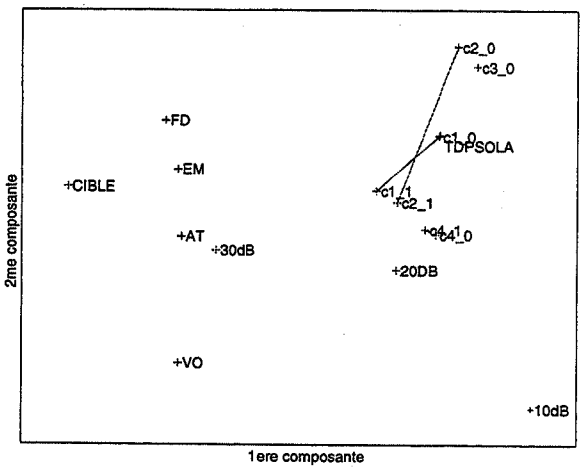


FIG. 3: Projection des divers codeurs sur le premier plan factoriel obtenu par ACP des distorsions moyennes qu'ils obtiennent sur les diverses tâches de transplantation. Les distorsions moyennes produites par l'ensemble des codeurs sur chaque tâche (FD, EM, AT, VO) ont été également projetées.

Mesures de distorsion Les nombreuses mesures de distorsion utilisées dans ces études estiment la distance ou la corrélation existant entre les représentations temps-fréquence des signaux comparés. Elles sont sensibles à diverses caractéristiques et il est impossible de sélectionner une seule et unique mesure capable de reproduire les jugements divers demandés aux sujets. De plus, ces mesures délivrent une mesure dynamique (voir Figure 2) qu'il est difficile de relier à une MOS globale (voir cependant [12]) : les mesures sont donc souvent moyennées.

Carte d'évaluation globale Au lieu de choisir une seule mesure, nous avons laissé à un algorithme de post-traitement le soin de "choisir" parmi un jeu de distorsions une combinaison permettant de différencier au mieux les résultats des codeurs. Suivant les propositions de Hansen et Pellom [11], nous avons calculé plusieurs mesures de distorsions centisecondes, entre cibles et transplantations prosodiques, utilisant : la pente spectrale pondérée (WSS), le quotient de similitude logarithmique (LLR) où les rapports d'aires logarithmiques (LAR). Cette liste n'est pas limitative et peut être amendée. Chaque codeur est ainsi caractérisé par a jeu de 90 distorsions moyennes (3 mesures x 15 tâches x 2 caractéristiques (moyenne, écart-type)). 10 systèmes ont été testés : TDPSOLA, 4 codeurs (c1_0, c2_0, c3_0, c4_0) ainsi que 3 versions améliorées (c1_1, c2_1, c4_1). c1, c2 et c4 sont des codeurs sinusoidaux alors que c3 utilise la prédiction linéaire (RELP). La cible bruitée à divers niveaux de bruit a été ajoutée (rapport signal sur bruit (RSB) de 10dB, 20dB et 30dB).

De manière à produire une représentation qui reflète la distance globale de chaque codeur à la cible et qui maximise leurs différences, ce jeu de 9x90 distorsions a été projeté sur le premier plan factoriel (voir Figure 3). Les trois premières composantes de l'analyse factorielle normalisée expliquent respectivement 87.7, 7.3 et 2.9 % de la variance des mesures de distorsion. Nous avons projeté de même les moyennes obtenues

par tous les codeurs sur chaque corpus, les points sont figurés par (AT,EM,VO,FD).

Commentaires Globalement les codeurs se situent à une dégradation équivalente de 10 à 20dB de RSB. Les versions améliorées des codeurs correspondent bien à un rapprochement de la projection du système vers la cible (de manière très nette pour c1_1 et c2_1 qui surpassent nettemant les performances de TDP-SOLA). L'évolution de la projection des signaux bruités et le placement des diverses projections des corpus sur le plan factoriel permettent d'interpréter le premier axe factoriel comme une mesure du RSB, le deuxième comme une distorsion au voisement - ce qui explique pourquoi un RSB de 10dB a une ordonnée plus forte que les autres RSB : les mesures de distorsion étant sensibles aux formants, lorsque ceux-ci sont noyés dans le bruit, la distorsion augmente très rapidement. De plus, les systèmes c2_0 et c3_0 ont des difficultés identifiées au niveau de la définition et du traitement des trames non-voisées. D'ailleurs, le corpus VO n'a pas été traité à nouveau lors de la deuxième soumission du codeur c2 : les résultats de c2_1 résultent bien en recentrage en ordonnée exprimant un réequilibrage des distorsions moyennes obtenues sur les segments voisés et non-voisés.

Conclusions

Il est facile de mettre à disposition des ressources sous format électronique permettant une évaluation cumulative et modulaires des systèmes de synthèse en vue d'un diagnostic aussi bien que d'une évaluation comparative. Nous espérons compléter ce serveur suivant trois directions : (1) augmenter l'éventail des descripteurs prosodiques des stimuli, (2) valoriser les propriétés des codeurs grâce à l'introduction de nouvelles tâches, notamment de concaténation de stimuli ou de manipulation de timbre, (3) mettre à disposition des méthodologies d'évaluation objective et cumuler les résultats d'évaluation subjective opérés sur ces stimuli. Nous espérons que la communauté francophone nous aidera dans cette démarche.

Remerciements

Ce travail a été supporté par le Cost258 et l'ARC-B3 de l'AUPELF-UREF. Grand merci à Eric Keller, Alex Monaghan, Eduardo Rodríguez Banga et Erhart Rank.

Bibliographie

[1] Almeida, L.B. and Silva, F. Variable-frequency synthesis : an improved harmonic coding scheme. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 27.5.1–4, 1984.

[2] Bailly, G., Barbe, T., and Wang, H. Automatic labelling of large prosodic databases : tools, methodology and links with a text-to-speech system. In Bailly, G. and Benoît, C., editors, *Talking Machines : Theories, Models and Designs*, pages 323–333. Elsevier B.V., 1992.

[3] Barbosa, P. *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France, 1994. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[4] Böeffard, O. and Violaro, F. Improving the robustness of text-to-speech synthesizers for large prosodic variations. In *ETRW on Speech Synthesis*, pages 111–114, New Paltz - New York, 1994.

[5] Campbell, W.N. Synthesizing spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing prosody : Computational models for processing spontaneous speech*, pages 165–186. Springer Verlag, 1997.

[6] Charpentier, F. and Moulines, E. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, 9(5-6) :453–467, 1990.

[7] Ding, W., Fujisawa, K., and Campbell, N. Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification. In *ETRW Workshop on Speech Synthesis*, pages 191–194, Jenolan Caves - Australia, 1998.

[8] Dutoit, T. High quality text-to-speech synthesis : A comparison of four candidate algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 565–568, Adelaide - Australia, 1994.

[9] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. The MBROLA project : towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the International Conference on Speech and Language Processing*, volume 3, pages 1393–1396, Philadelphia - USA, 1996.

[10] Gobl, C. and Chasaide, N. Acoustic characteristics of voice quality. *Speech Communication*, 11(4-5) :481–490, 1992.

[11] Hansen, J.H. and Pellom, B.L. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the International Conference on Speech and Language Processing*, volume 6, pages 2819–2822, 1998.

[12] Hansen, M. and Kollmeier, B. Continuous assessment of time-varying speech quality. *Journal of the Acoustical Society of America*, 105(5) :2888–2899, 1999.

[13] Hermes, D.J. Synthesis of breathy vowels : Some research methods. *Speech Communication*, 10 :497–502, 1991.

[14] Klabbers, E. and Veldhuis, R. On the reduction of concatenation artefacts in diphone synthesis. In *Proceedings of the International Conference on Speech and Language Processing*, volume 5, pages 1983–1986, 1998.

[15] Quatieri, T.F. and McAulay, R.J. Phase coherence in speech reconstruction for enhancement and coding applications. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 207–210, 1989.

[16] Sonntag, G.P., Portele, T., Haas, F., and Köhler, J. Comparative evaluation of six German TTS systems. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 251–254, Budapest, 1999.

[17] Stylianou, Y. Concatenative speech synthesis using a Harmonic plus Noise Model. In *ESCA/COCOSDA Workshop on Speech Synthesis*, pages 261–266, Jenolan Caves, Australia, november 1998.

[18] Syrdal, A.K., Möhler, G., Dusterhoff, K., Conkie, A., and Black, A.W. Three methods of intonation modeling. In *ESCA/COCOSDA Workshop on Speech Synthesis*, pages 305–310, Jenolan Caves, Australia, november 1998.

[19] Taylor, P., Black, A.W., and Caley, R. The architecture of the FESTIVAL speech synthesis system. In *ESCA/COCOSDA Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia, 1998.

[20] Veldhuis, R. and Hé, H. Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communication*, 18 :257–279, 1996.

[21] Yvon, F., de Mareuil, P.B., d'Alessandro, C., Aubergé, V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Glodman, J.F., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., and Zellner, B. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. *Computer Speech and Language*, 12 :393–410, 1998.

Génération de la prosodie par superposition de contours chevauchants: application à l'énonciation de formules mathématiques

Bleicke Holm & Gérard Bailly

Institut de la Communication Parlée - UMR CNRS $n°5009$ - INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél: ++33 04 76 57 48 50 - Fax: ++33 04 76 57 47 10
e-mail: holm@icp.inpg.fr - http://www.icp.inpg.fr/

Abstract

We present here a model for generating prosody by superposing overlaping multiparametric contours. These contours are associated with high-level communication tasks such as segmentation, hierarchisation or emphasis of units. We propose a scheme for automatically learning these contours and apply this new paradigm to the enunciation of mathematical formulae.

1. Introduction

La prosodie participe à la transmission d'information linguistique dans l'acte de parole. Parmi les fonctions linguistiques assumées les plus étudiées, on trouve la fonction de hiérarchisation/segmentation d'unités. Cutler note ainsi : "*Two of the major uses of prosodic information in situations of communication are to encode salience and segmentation.*" [6, p.264]. Si tout le monde s'accorde maintenant sur le rôle fondamental joué par la prosodie dans l'acquisition [14] et la structuration du langage, de grosses divergences apparaissent sur les moyens d'encodage de l'information. Dans le cadre d'un modèle de génération automatique de la prosodie, il s'agit de fournir une évolution acceptable de paramètres prosodiques en fonction de la structuration de l'énoncé. Se pose ainsi la double question du Quoi et du Comment : quelles sont les informations véhiculées et comment sont-elles encodées ?

L'encodage d'informations linguistiques, supposées discrètes et dénombrables, par un continuum a été décrit par de multiples approches. La plupart utilisent une interface phonologique autonome où la structuration linguistique est "traduite" en une structuration phonologique organisant les faits prosodiques saillants en des structures tonales et accentuelles. A cette étape de transfert phonologique, doit alors succéder une étape de génération phonétique où les structures phonologiques sont traduites en continuums prosodiques.

Les systèmes de génération automatique diffèrent dans la richesse des structures phonologiques exploitées : ainsi, Black et Hunt [5] utilisent la transcription ToBI [17], Di Cristo et al IntSyn [10]. D'autres auteurs utilisent une description plus économique de la structure phonologique en termes de groupes accentuels [18, 12, 13]. Ce dernier groupe d'approches est très hétérogène : Malfrère et al [12] portent leur effort sur une concaténation élégante de contours faiblement étiquetés par rapport à leur fonction alors que Pasdeloup [16] focalise sur un bon découpage en groupes accentuels.

Les modèles phonétiques utilisés lors de l'étape de génération vont de la simple connection de points cibles pour la génération de la mélodie [10] à des modèles plus complexes. La notion de superposition est largement répandue au moins ce qui concerne deux composants qui sont distingués : phrase et groupe accentuel. La superposition peut être très explicite, comme dans [13] ou en forme de l'imposition d'une ligne de déclinaison sur laquelle s'ajoutent les contributions dues aux accents. Même [5] incorpore une superposition car la réalisation des accents est exprimée par rapport à une fréquence de référence qui varie au cours de la phrase. Chez Traber [18], l'usage d'une large fenêtre sur l'énoncé et de réseaux récurrents permet de compenser la pauvreté de la description phonologique d'entrée : le réseau peut "calculer" de manière implicite des niveaux de représentation intermédiaire.

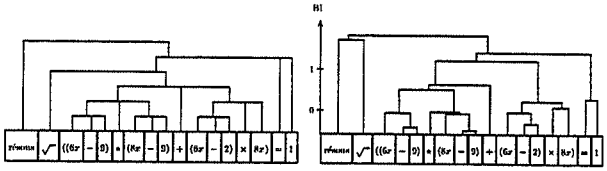
Notons finalement la proposition d'Aubergé [1] qui contraste avec cette décomposition des tâches : c'est le niveau d'encodage phonétique qui traduit "directement" la structuration linguistique par superposition de contours globaux coextensifs aux unités qu'ils qualifient. Dans cet article, nous présentons un encodage à la fois *simple* par le mécanisme d'encodage qu'il utilise et *riche* par les capacités de structuration et de coarticulation qu'il offre. L'encodage opère par superposition de contours chevauchants. Cet encodage est *simple* et général car chaque contour code une fonction de hiérarchisation/segmentation d'unités quelque soit cette unité et sa hauteur dans la hiérarchie linguistique. *Riche* parce que le nombre de niveaux superposés et donc le degré de hiérarchisation de l'énoncé est illimité.

2. L'énonciation de formules mathématiques

Notre corpus a été conçu afin de pouvoir étudier la manière avec laquelle la prosodie peut encoder des relations de dépendance à l'intérieur d'une phrase. Nous avons choisi des formules mathématiques lues (FM) parce qu'on y trouve une structure syntaxique profonde qui est souvent ambiguë à l'oral et qui nécessite un recours renforcé à la structuration de l'énoncé par la prosodie. Nous avons restreint le domaine des

formules aux équations algébriques issues de l'enseignement de niveau 3^e. Le corpus a été généré automatiquement en variant systématiquement la profondeur syntaxique et la longueur syllabique des constituants pour couvrir statistiquement l'influence de ces paramètres sur la stratégie prosodique – tout en gardant une taille de corpus raisonnable. Nous obtenons un corpus de 157 FM[1]. Il est lu par un locuteur français à qui nous avons donné la consigne de ne pas utiliser d'indicateurs lexicaux de structure comme "ouvrez les parenthèses".

Ainsi que montré sur la figure 1 et plus amplement commenté dans [11], la structuration prosodique de formules énoncées est riche et profonde. De plus, chaque ajout d'un niveau syntaxique supplémentaire se traduit par un accroissement de la hauteur de la structure de performance correspondante, ce qui est bien en adéquation avec le modèle de superposition dont la description suit.



(a) Syntaxe (b) Performance

Fig. 1: Structure syntaxique et structure de performance [8] de l'énoncé d'une formule.

3. Le modèle

Dans la proposition initiale d'Aubergé [1], chaque niveau linguistique génère en fonction de sa nature, de sa fonction et de sa longueur un ensemble de contours multiparamétriques qui seront ensuite concaténés et superposés. Morlec [15] a traité le problème de l'apprentissage des contours, de leur expansion avec la taille des unités et de leur enchaînement en proposant une architecture connexionniste modulaire. Cette approche a été validée aussi bien en analyse qu'en synthèse.

Ce modèle initial et son implémentation souffrent cependant de deux limitations : (a) le modèle suit de trop près la syntaxe au sens où un contour est indexé par des caractéristiques syntaxiques souvent redondantes et (b) si l'implémentation par réseau récurrent gère mieux la coarticulation que la simple concaténation initiale – où l'étiquetage en sous-groupes était conditionné à des contraintes phonotactiques –, elle fait perdre la notion initiale de lexique de contours attachés à une fonction linguistique.

Le modèle que nous développons ici vise à résoudre ces deux problèmes par (a) l'adoption d'une hypothèse additionnelle donnant au modèle morphologique initial une véritable morphophonologie fonctionnelle, et (b) l'adoption d'un modèle superpositionnel de contours *chevauchants*.

3.1. *Une morphophonologie fonctionnelle*

La morphophonologie fonctionnelle que nous proposons suppose que l'intonation met au service de la syntaxe des fonctions générales, que l'intonation partage avec la syntaxe mais qu'elles assument de manière autonome et donc pas forcément de manière congruente. Parmi ces fonctions plus autonomes, on peut citer les fonctions attitudinales [15] ou émotionnelles que l'on peut caractériser par des formes prosodiques prototypiques. Parmi les fonctions plus synergétiques, on peut citer la fonction de segmentation/hiérarchisation ou de mise en relief : nous allons ici poser et tester l'hypothèse que la prosodie possède les moyens de *segmenter/hiérarchiser* des unités consécutives en centrant un contour sur la jonction entre ces deux unités (voir figure 2(en haut)). La hiérarchie est alors encodée par la superposition de contours qui se chevauchent et qui ont des empans différents mais limités aux unités adjacentes à la jonction mise en valeur.

L'encodage par contours permet de distribuer l'information au cours du temps en ne limitant pas l'accès à la fonction au décodage d'éléments saillants du discours. Cet encodage distribué permet d'expliquer les phénomènes d'anticipation [2], d'attente et de prédiction sur la suite d'un énoncé [9]. Il semble cependant évident que cet encodage doive être programmé sur un empan temporel limité.

3.2. *Apprentissage des contours*

L'inversion d'un tel modèle est un problème mal-posé : comment obtenir les contours localisés à partir d'un corpus qui ne contiendra jamais ces contours isolés mais toujours superposés ? La stratégie adoptée par Aubergé [1], reprise par Morlec [15], consiste à apprendre hiérarchiquement les contours, de celui possédant l'empan le plus large jusqu'au plus petit. A la différence de Morlec où chaque module génère des contours *globaux* associés à un niveau hiérarchique *fixe*, chaque module ici prend en charge la génération de tout contour localisé assumant la fonction désirée : le module génère des contours ancrés sur une partie du discours et non plus sur le discours entier. Il est appliqué plusieurs fois aussi bien de manière séquentielle (des fonctions identiques apparaissant à divers points du discours) que hiérarchique (appliqué à des niveaux linguistiques différents pour par exemple segmenter/hiérarchiser des unités de différentes tailles). La décomposition du contour global du discours en contours localisés et ancrés ne peut donc suivre l'apprentissage hiérarchique.

Nous proposons un apprentissage itératif permettant à chaque module de s'approprier petit à petit des parties du contour global. Chaque cycle d'apprentissage comporte deux phases : (1a) calcul de la différence entre contours globaux observés et prédits, (1b) calcul des cibles (contours localisés) de chaque module et, (2) apprentissage des cibles par les modules. (1a) est triviale – sauf au premier cycle où nous utilisons une prédiction égale à zéro. La base pour le calcul des nouveaux cibles en (1b) sont les contours localisés prédits. On y ajoute de manière uniforme une fraction

[1] Pour plus de précisions sur le corpus on peut se référer à [11].

de la différence calculée en (1a) : pour chaque trame[2], chaque contour localisé contribuant au contour prosodique de cette trame reçoit la différence en ce point divisée par le nombre total de contours localisés contributeurs – en s'assurant ainsi que la superposition des cibles est égale au contour observé. La phase (2) s'appuie sur l'implémentation des modules en forme de *réseaux neuronaux* : ils sont bien adaptés pour mettre en cohérence les nombreux exemplaires de réalisation d'un même contour localisé.

Chaque module[3] apprend à associer les vecteurs d'entrée aux cibles. L'information fournie en entrée est très simple : la distance (en nombre de GIPC et en pourcent) de la trame courante du début de l'unité, du point d'ancrage et de la fin de l'unité. Ainsi, *une* fonction linguistique est encodée par *un* module qui gère l'évolution des contours en fonction de l'information phonotactique.

3.3. L'énonciation de formules

La fonction de segmentation/hiérarchisation est la plus prégnante des fonctions sollicitées lors de l'énonciation de formules. La fonction de hiérarchisation est essentiellement sollicitée pour mettre en valeur l'enchâssement des opérateurs. Les opérateurs présents dans notre corpus sont simples : on distingue des opérateurs unaires (comme $\sqrt{\ldots}$) et des opérateurs binaires (comme +). Nous travaillerons prochainement sur des opérateurs plus complexes du type $\sum$.

Ainsi selon l'hypothèse que l'intonation suit ici la hiérarchie syntaxique, nous nous contraignons à distinguer dans notre corpus que deux types de dépendance : dépendance gauche (G) – entre un opérateur et son opérande[4] à gauche et, dépendance droite (D) entre un opérateur et son opérande à droite. A chaque frontière, entre deux feuilles de l'arbre syntaxique, nous posons un marqueur G ou D. Ils symbolisent chacun une relation de dépendance, à l'image des systèmes de marques syntaxiques proposées dans la littérature [7, 3]. Deux modules prendront dès lors en charge la génération de toutes les instances de ces marques en fonction des empans correspondants.

Les relations de présuppositions entre phrases ne sont pas factorisées avec les précédentes : nous avons ainsi un troisième module (M) prenant en charge le contour ancré à la frontière entre "Résous :" et la formule. Son empan sera l'énoncé entier.

4. Premières observations

Afin de donner une idée de la qualité des contours générés, nous présentons dans la figure 2 pour un exemple la prédiction (trait plein) avec le contour observé (pointillés). On observe un comportement similaire mais, notamment l'accent sur la sixième syllabe n'est pas bien rendu.

[2]L'unité de programmation prosodique est pour l'instant le GIPC [4] caractérisé par trois points de F0 et un coefficient d'allongement.

[3]Ici : M, D et G – voir paragraphe 3.3

[4]Par "opérande" nous entendons l'ensemble des feuilles appartenant à un sous-arbre en direct voisinage de l'opérateur.

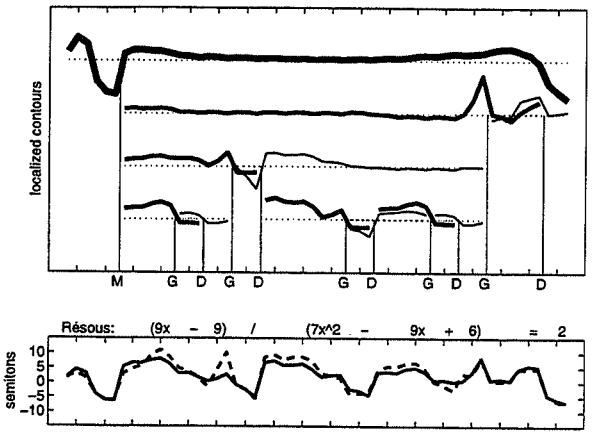


Fig. 2: En haut : contours localisés et ancrés de f_0. On note le point d'ancrage de chaque contour dont on distingue trois types : hiérarchie de phrase (M), dépendance gauche (G) et dépendance droite (D). Les pointillés indiquent le zéro pour chaque contour. — En bas : contour observé (tirets) contour prédit par superposition (plein).

Il est indispensable d'analyser le comportement des contours localisés générés. Regardons quelques contours du coefficient d'allongement que produit le module G (figure 3). Dans notre corpus, ces contours apparaissent toujours pour un opérateur et son opérande gauche. Dans la ligne en bas, on voit que la frontière est marquée par un important allongement[5] du GIPC avant l'opérateur. L'allongement décroît avec la longueur de l'unité gauche (en remontant les lignes dans la figure 3). La première ligne des contours échappe à cette régularité : un opérande d'une syllabe est clairement marqué par un fort allongement.

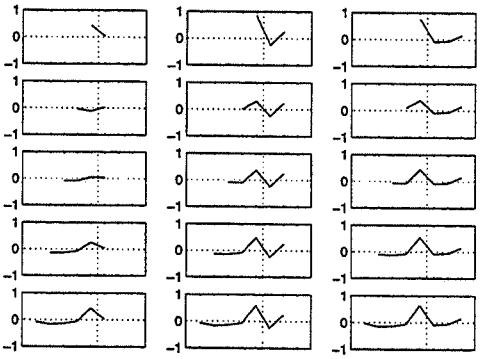


Fig. 3: Sortie (coefficient d'allongement) du module G pour différentes longueurs des unités : de 1 à 5 GIPC avant le point d'ancrage (indiqué par les lignes verticales) et de 1 à 3 après celui-ci.

5. Évaluation

En guise de première évaluation, nous présentons une comparaison des RMS-erreurs pour plusieurs méthodes de génération des contours : (1) "global–non-

[5]Voire une pause. Pour le traitement des pauses dans notre modèle, on peut se référer à [4].

apprentissage	C1+C2	C1	
test	C1+C2	C1	C2
méth.(1)	0.205±0.010	0.185±0.005	0.248±0.003
méth.(2)	0.217±0.004	0.206±0.003	0.248±0.010
méth.(3)	0.213±0.003	0.208±0.002	0.238±0.006
méth.(1)	1.96±0.03	1.81±0.04	2.40±0.06
méth.(2)	2.07±0.01	1.98±0.02	2.34±0.04
méth.(3)	2.13±0.02	2.11±0.03	2.31±0.02

Tab. 1: Erreur de prédiction et son écart-type pour plusieurs méthodes et différents corpus d'apprentissage. En haut : coefficient d'allongement, en bas : f_0 en semitons.

hiérarchique", (2) "global–hiérarchique" et (3) "local–itératif". Les deux premières méthodes suivent l'architecture de Morlec, mais diffèrent par la gestion de l'apprentissage : dans (1) l'apprentissage des modules se fait simultanément (s'approchant ainsi du modèle de Traber) tandis que dans (2) l'apprentissage est hiérarchique (comme proposé par Morlec). (3) correspond au modèle proposé dans cet article. Dans le tableau 1 la première colonne correspond à un apprentissage sur le corpus total (C1+C2). Les deux autres montrent les erreurs pour des apprentissages sur une moitié du corpus (C1) – au milieu : l'erreur sur le corpus d'apprentissage C1, et à gauche : celle sur celui de test (C2). Chaque valeur est issue de 4 apprentissages du même type. Bien que l'erreur totale augmente avec le numéro de version, on observe une amélioration pour la généralisation (colonne de droite) – ce qui est le critère plus important pour l'évaluation du modèle.

Notons finalement que pour le nouveau modèle, le nombre de paramètres à estimer lors de l'apprentissage des modules chute d'un facteur 2 par rapport aux concurrents.

6. Conclusions et perspectives

L'évolution du modèle que nous avons présenté ici évite certaines lacunes de la version précédente : (1) Il n'était pas possible d'exploiter aisément la récursivité des relations de dépendance. Il nécessitait notamment un changement d'architecture (multiplication des modules) selon les extensions que l'on voudrait apporter aux structures syntaxiques qu'il est capable de traiter. Dans la nouvelle version, une fois les relations de dépendance identifiées, le nombre de niveaux de récursivité reste flexible et exerce une influence sur l'apprentissage non sur l'architecture. On peut même espérer obtenir une génération de contours acceptable pour une structure plus profonde que celle du corpus d'apprentissage. (2) La relation entre fonction linguistique et génération de contour est maintenant plus étroite : désormais *un* seul module de génération encode *une* fonction linguistique. (3) Les contours sont localisés là où ils sont censés d'agir : aux unités qu'ils mettent en relation.

Néanmoins, nous ne disposons que d'une première implémentation du modèle. Mis à part la comparaison numérique des contours prédits et observés et l'analyse de la cohérence des familles de contours, le test crucial devra être une évaluation perceptive de la prosodie générée.

Bibliographie

[1] Aubergé, V. *La synthèse de la parole : "des règles aux lexiques"*. PhD thesis, Université Pierre Mendès-France, Grenoble – France, 1991. sous la direction de J. Rouault.

[2] Aubergé, V., Grépillat, T., and Rilliard, A. Can we perceive attitudes before the end of sentences ? The gating paradigm for prosodic contours. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 871–874, Rhodes - Greece, 1997.

[3] Bailly, G. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8 :137–146, 1989.

[4] Barbosa, P. *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France, 1994. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[5] Black, A.W. and Hunt, A.J. Generating f0 contours from tobi labels using linear regression. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1385–1388, 1996.

[6] Cutler, A. and Norris, D. Prosody in situations of communication : salience and segmentation. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, pages 264–270, Aix-en-Provence, France, 1991.

[7] Emerard, F. Synthèse par diphones et traitement de la prosodie. Thèse de troisième cycle, Université de Grenoble III, Grenoble, France, 1977.

[8] Gee, J.P. and Grosjean, F. Performance structures : a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15 :411–458, 1983.

[9] Grosjean, F. How long is the sentence ? prediction and prosody in the on-line processing of language. *Linguistica*, 21 :501–529, 1983.

[10] Hirst, D.J. and Di Cristo, A. *Intonation systems : a survey of twenty languages.* Cambridge University Press, Cambridge, 1998.

[11] Holm, B., Bailly, G., and Laborde, C. Performance structures of mathematical formulae. In *Proceedings of the International Congress of Phonetic Sciences*, volume 2, pages 1297–1300, San Francisco, USA, 1999.

[12] Malfrère, F., Dutoit, T., and Mertens, P. Automatic prosody generation using suprasegmental unit selection. In *ESCA/COCOSDA Workshop on Speech Synthesis*, pages 323–328, 1998.

[13] Möbius, B., Pätzold, M., and Hess, W. Analysis and synthesis of german f0 contours by means of fujisaki's model. *Speech Communication*, 13 :53–61, 1993.

[14] Morgan, J.L. and Demuth, K. *Signal to syntax : an overview.* Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1996.

[15] Morlec, Y. *Génération multiparamétrique de la prosodie du français par apprentissage automatique.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble – France, 1997. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[16] Pasdeloup, V. A prosodic model for french text-to-speech synthesis : A psycholinguistic approach. In Bailly, G., Benoît, C., and Sawallis, T., editors, *Talking Machines : Theories, Models and Designs*, pages 335–348. Esevier B.V., 1992.

[17] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. Tobi : a standard for labeling english prosody. *Proceedings of the International Conference on Speech and Language Processing*, 2 :867–870, 1992.

[18] Traber, C. Fo generation with a database of natural fo patterns and with a neural network. In Bailly, G. and Benoît, C., editors, *Talking Machines : Theories, Models and Designs*, pages 287–304. Elsevier B.V., 1992.

Ressource standard pour le français : un large lexique orthographique-phonétique

François Yvon, Christophe d'Alessandro, Véronique Aubergé, Philippe Boula de Mareüil, Jacqueline Vaissière

ENST - Dpt. Informatique
46 rue Barrault
F-75634 Paris cedex 13
yvon@inf.enst.fr

LIMSI-CNRS
BP 133
F-91403 Orsay cedex
{cda,mareuil}@limsi.fr

ICP - Univ. Stendhal
1180 avenue Centrale - BP25
F-38040 Grenoble cedex 9
auberge@icp.inpg.fr

ILPGA - Univ. Paris III
19 rue des Bernardins
F-75 005 Paris
jvaiss@msh-paris.fr

RÉSUMÉ

This paper reports on a project aiming at the semi-automatic development of a large orthographic-phonetic lexicon for French, based on Multext lexicon. It details the various stages of the project, with an emphasis on the methodological and design aspects. Information regarding the lexicon's content is also given, together with a description of tools which should facilitate its exploitation.

1. INTRODUCTION

Cet article décrit le développement semi-automatique d'un lexique phonétisé de large couverture pour le français. L'objectif principal de ce projet était de capitaliser les données et l'expertise disponibles en matière de transcription automatique du français, pour développer des bases de données textuelles phonétisées pouvant servir de ressources pour le développement ou l'évaluation d'applications (en particulier de traitement de la parole). Les campagnes d'évaluation des systèmes de traitement de la parole organisées sous l'égide de l'AUPELF-UREF ont clairement mis en évidence l'importance de la libre disponibilité de telles ressources pour le français. Dans le cadre de ce projet, l'effort a été mis sur le développement de ressources génériques de type dictionnairique : à partir des entrées du lexique Multext [IV94], standardisé pour un niveau de langue générale, ont été dérivées des entrées phonétiques, référentielles et variantes. À ces entrées ont été ajoutées des listes orthographique-phonétiques de noms propres et d'acronymes. Ce lexique se destine à être un complément des données existant pour le traitement automatique, comme BDLEX [PdCFP92], et une alternative disponible librement aux dictionnaires commerciaux grand public tels que Le Robert.

L'originalité principale de ce travail provient de l'utilisation de systèmes de transcription automatique développés dans trois laboratoires. L´évaluation des systèmes de transcription du français [YBd+98] a en effet montré que, même si le problème de la transcription automatique reste difficile, la précision des systèmes actuels les rend capables de produire des phonétisations acceptables pour la majorité des mots du lexique commun. En nous appuyant sur ces systèmes, nous avons donc pu mettre en place un processus de production semi-automatique des transcriptions phonétiques, permettant d'accélérer le développement des transcriptions, et de parvenir à un lexique de large couverture en un temps réduit.

Une préoccupation majeure a présidé à la conception et au développement de ce lexique, celle de la **réutilisabilité** des ressources. Cela s'est traduit dans notre contexte par trois contraintes :

- la prise en compte des besoins de tous les types d'utilisations potentielles de ces ressources (reconnaissance et synthèse de la parole, correction orthographique, apprentissage des langues...). La question de la représentation de la variabilité de la prononciation, dont la modélisation est nécessaire en particulier pour les applications de reconnaissance de la parole, a ainsi fait l'objet d'une réflexion approfondie ;
- l'adhésion aux standards de faits disponibles ou émergents, aussi bien en matière de formats, qu'en matière de représentations des informations linguistiques ;
- l'anticipation des extensions possibles de ces ressources.

Ce projet s'est déroulé en trois étapes principales, qui sont détaillées dans les sections qui suivent : une étape de conception du lexique, incluant une réfléxion méthodologique et des spécifications précises de la nature et du format des données à produire ; une étape de production des transcriptions phonétiques ; enfin une étape consacrée au développement d'outils d'interface permettant l'exploitation des transcriptions produites.

2. CONCEPTION DU LEXIQUE

2.1. Lexique original

Le lexique de base est une des versions du lexique du français développé dans le cadre du projet Multext [IV94]. Cette version est celle du projet GRACE portant sur l'évaluation des systèmes d'étiquetage morpho-syntaxique du français [ABM+95]. Elle contient 310 332 formes orthographiques fléchies, soit 27 873 lemmes différents. Chaque entrée du lexique original contient trois champs, correspondant respectivement à la forme graphique, à la forme graphique du lemme associé, et à la description morpho-syntaxique de la forme. Ce dernier champ spécifie la catégorie syntaxique principale, ainsi que diverses informations complémentaires pertinentes pour cette catégorie (par exemple, genre et nombre pour les formes nominales ; temps,

mode, nombre et personne pour les formes verbales ; etc.) [CM96, RLP97]. Ce lexique présente ainsi l'avantage de contenir des informations morpho-syntaxiques riches, exprimées dans un jeu d'étiquettes standard, utilisé dans d'autres projets en traitement automatique de la langue. De plus, les informations associées à une forme permettent à la fois d'effectuer la désambiguïsation des homographes-hétérophones, et de retrouver la forme étendue pour les abréviations. La table 1 contient un échantillon d'entrées lexicales.

TAB. 1 – *Exemples d'entrées lexicales*

Forme	Lemme	Étiquette
couvent	couvent	`Ncms`
couvent	couver	`Vmip3p-`
mat	mat	`Afpms`
mat	mat	`Ncms`
revenir	revenir	`Vmn--`
absolu	absolu	`Afpms`
MM.	messieurs	`Ncmp`
messieurs	monsieur	`Ncmp`

Nous avons, dans un deuxième temps, étendu ce lexique, par addition d'une dizaine de milliers de noms propres (noms de lieux, de personnes, de sociétés et d'institutions). Ces noms propres sont, pour l'essentiel, issus de l'analyse automatique de corpus journalistiques. Ces extractions ont été réalisées grâce au système d´étiquetage morpho-syntaxique développé au LIA (Avignon) [SBEB+96]. Ces listes de noms propres ont été vérifiées manuellement, et, le cas échéant, réétiquetées. À cet effet, nous avons étendu marginalement le jeu d´étiquettes Multext, qui ne distingue, à l'aide du trait "semantic type", que deux sortes de noms propres : les noms de ville et de pays (*Londres*, catégorisé `Np-s-c`), et les noms de société (*IBM*, catégorisé `Npms-s`). Ont été notamment ajoutées les nouvelles valeurs suivantes pour la catégorisation sémantique : `f`, pour les prénoms (*Jean*, catégorisé `Npms-f`), et l, pour les noms de personnes (*Dupont*, catégorisé `Np-s-l`). Ces distinctions supplémentaires ont également été effectuées dans le lexique Multext original.

2.2. *Spécification de la transcription phonétique*

La première étape du travail de spécification a consisté à préciser le niveau de description linguistique encodé dans les transcriptions. Compte-tenu des contraintes (utiliser des systèmes de transcription déjà existants) et des pratiques en la matière, nous avons opté pour une description normative (correspondant à la norme « parisienne »), sur la base d'une transcription phonétique « large ». Pour noter les transcriptions, nous avons adopté l'alphabet SAMPA [1] [GMe97], qui constitue une référence bien établie.

Le développement d'un lexique phonétique pose ensuite le problème de la description de la *variabilité* : chaque forme orthographique peut se réaliser oralement de multiples façons différentes. Nous avons choisi de représenter cette variabilité de manière intensionnelle, en adoptant un système de représentation permettant de représenter, dans un champ unique, l'ensemble des phonétisations possibles d'une forme. Cette approche, qui est adoptée dans divers lexiques de prononciation (tels que BDLEX [PdCFP92]), nous a paru plus conforme avec l'esprit du projet Multext, que l'alternative consistant à lister explicitement les différentes variantes, et qui est mise en œuvre, par exemple, dans le lexique CELEX [Bur90]. Sa mise en œuvre demande toutefois de concevoir un format permettant de représenter la variabilité, et de développer des procédures permettant d'engendrer les différentes variantes possibles.

À la suite en particulier de [Lap89, LD90], on peut en fait distinguer, au niveau phonologique, plusieurs types de variabilité :

- une variabilité systémique, conditionnée par l'environnement linguistique. Ce type de variabilité est illustré par les trois réalisations différentes du mot *six* dans les occurrences suivantes (exemple tiré de [Lap89]) :
  - Luc a six (/si/) billes ;
  - Luc a six (/siz/) ans ;
  - Luc en a six (/sis/) ;
- une variabilité contextuelle ou stylistique, c'est-à-dire conditionnée par la réalisation des unités. Entrent dans ce cadre aussi bien les variantes idiolectales ou sociolectales, que les variations entraînées par la prosodie, liées au contexte d'élocution et de coarticulation, qui entraînent par exemple des variations dans la réalisation optionnelle des liaisons, des diérèses, des e-muets, et des assimilations.

Ces considérations nous ont amenés à enrichir SAMPA d'une part par un ensemble des diacritiques, d'autre part par des méta-caractères permettant de noter de manière compacte les groupes variables. Ces symboles sont présentés dans la table 2.

TAB. 2 – *Symboles ajoutés à SAMPA*

Diacritiques		Méta-caractères	
-	optionnel (phonétique)	(	début de groupe
.	voisement	)	fin de groupe
0	dévoisement	{	début de finale
<	ouverture	}	fin de finale
>	fermeture	|	alternative
_	phonème latent		

Les diacritiques permettent de noter en particulier les phonèmes «optionnels» (géminées, e-muets élidables, consonnes finales...) ; les neutralisations de timbres, les assimilations de voisement. À l'instar du symbole de nasalisation ~, ces diacritiques spécifient la prononciation du symbole qui les précède. Alors que BDLEX utilise des archiphonèmes pour les voyelles intermédiaires quand l'opposition de timbre est neutralisée, les diacritiques < et > permettent ici d'évaluer simplement la précision d'une transcription dans un alphabet phonétique traditionnel. Il en va de même pour le 'e' caduc (/@-/) et pour le comportement phonologique des finales. Les paires de symboles { } et () identifient les groupes dont la prononciation est variable : {...|...} sert à décrire les variantes de pronon-

1. Voir `http://www.phon.ucl.ac.uk/home/sampa/home.htm`.

ciation résultant respectivement de la non-réalisation ou réalisation d'une liaison ; (...|...) est utilisé pour les groupes à prononciation variable, la première alternative notant la prononciation préférée de la forme. La table 3 contient un ensemble de transcriptions illustrant l'utilisation de ces symboles. L'ensemble des

TAB. 3 – *Exemples de transcriptions phonétiques*

samedi	sam@-di	anecdote	anEk.dOt@-
vendredi	va~dr@di	absolu	ab0sO>ly
quatre	katr@-	ananas	anana(|s)
petite	p@-tit@-	grand	gRa~{|t_}
illégal	il-legal	divin	div{e~|in}
annoté	an-nO>te	six	si{s-|z_}
était	e<tE{|t_}	plus	ply{s-|z_}

consignes spécifiant la représentation des transcriptions, et édictant un certain nombre de principes permettant de déterminer la phonétisation d'une forme a été rassemblé dans un guide détaillé [dM98]. Ce guide, qui a permis à la fois de rendre cohérents les trois systèmes de transcription et d'orienter le travail de contrôle des transcriptions sera joint à la distribution finale du projet.

2.3. *Transcription automatique*

Les systèmes de transcription automatique disponibles à l'ICP, au LIMSI, et à l'ENST ont, dans un premier temps, été adaptés pour accomoder le format du lexique original et pour produire des phonétisations conformes aux spécifications adoptées. Des informations détaillées sur l'architecture de ces systèmes sont données dans [Aub91, dM97, Yvo96]. Chaque système a été utilisé pour produire indépendamment une phonétisation pour chaque entrée de Multext. Les trois transcriptions ainsi obtenues ont ensuite été alignées, de manière à identifier et à évaluer les cas de désaccord entre les laboratoires. Nous avons pour cela utilisé un algorithme implantant un calcul de distance d'édition tolérante entre chaînes, basé sur une grammaire pondérée de correspondances graphoniques.

Une première liste d'environ 35 000 formes a été corrigée manuellement, correspondant aux cas de désaccord entre les trois systèmes. Ce travail a nécessité l'intervention de trois experts pendant plusieurs mois, et le développement d'une interface ergonomique dédiée à cette tâche. Dans un second temps, les cas de désaccords entre deux laboratoires ont été analysés, donnant lieu à de nouvelles vagues de corrections, portant sur environ 12 000 formes. Au total, on peut considérer qu'environ 85% des transcriptions ont été automatiquement validées en croisant les résultats des différents systèmes ; les 15% restant faisant l'objet d'une expertise humaine.

De manière indépendante, des listes de noms propres ont été collectées selon la procédure décrite à la section 2.1, transcrites automatiquement à l'aide du système de l'ENST, et intégralement vérifiées.

3. EXPLOITATION DU LEXIQUE

3.1. *Formats*

Le lexique Multext est représenté informatiquement sous la forme d'un fichier texte, chaque champ étant séparé par une tabulation. Nous avons adopté un second format de représentation, plus approprié en particulier pour la visualisation, à savoir XML. Nous avons donc spécifié une DTD XML permettant de représenter de manière structurée les informations lexicales. Cette opération nous a conduits à toucher du doigt les limites de la représentation «plate», qui, en particulier, ne permet pas d'associer directement une forme au lemme correspondant, puisque la seule correspondance enregistrée dans Multext lie une forme orthographique à la *représentation orthographique* du lemme. Cette correspondance a été rétablie en associant à chaque entrée une clé numérique unique, et en recalculant les associations entre formes et lemmes.

3.2. *Visualisation*

Une interface permettant la consultation du lexique à travers le réseau Internet a été développée. Cette interface se fonde sur la construction à la demande de pages HTML permettant de visualiser le lexique XML, et inclut notamment des fonctionnalités de recherche d'une forme, d'un lemme ou d'une transcription. Il est également possible, par le jeu de liens hypertextes, d'accéder directement au lemme associé à une forme donnée. Cette interface sera distribuée avec l'ensemble des ressources.

3.3. *Outils de traitement*

Il est possible, à partir de la représentation des transcriptions phonétiques, de dériver un grand nombre d'informations complémentaires, qui sont utiles dans différents contextes. Ainsi par exemple, une fonction permet d'engendrer explicitement les différentes variantes primaires d'une forme. Ces variantes sont exprimées dans l'alphabet SAMPA «de base». Ainsi *annoté*, transcrit /an-nO>te/, donne lieu à quatre variantes primaires : /anOte/, /annOte/, /annOte/ et /annote/. Il est ensuite possible, pour chacune de ces transcriptions, de calculer des informations complémentaires, telles que la décomposition en syllabes, le nombre de syllabes, ou encore un squelette en cohortes CV... Il est également possible, en appliquant un jeu paramétrable de règles de récriture décrivant les principaux phénomènes d'assimilation, d'engendrer des variantes de prononciation supplémentaires, correspondant à des réalisation plus «relâchées» de la forme. Ainsi, par exemple : /mEdse~/ puis /mEtse~/ pour la forme *médesin*. L'ensemble de ces calculs est réalisé par un ensemble de fonctions écrites en Perl, qui sont distribuées avec le lexique.

4. CONCLUSION ET PERSPECTIVES

Un des apports principaux de ce projet a été de démontrer la viabilité de l'approche semi-automatique pour le développement de ressources linguistiques. Il est en particulier apparu que le recoupement des transcriptions produits par différents systèmes permettait

de détecter les erreurs de phonétisation avec une grande précision. Pour les laboratoires impliqués dans le projet, une retombée importante de ce projet a été une évaluation informelle des performances de leur système face à des données dictionnairiques, permettant de pointer un certain nombre de lacunes.

Ce projet a permis la création d'un lexique de prononciation, couvrant de manière assez exhaustive la langue générale et intégrant un nombre significatif de noms propres et de sigles. Ce lexique, fondé sur les ressources Multext, sera distribué sous la forme d'une base de donnée XML. Il devrait permettre de faciliter le développement et l'évaluation de systèmes de transcription graphème-phonème pour des applications du traitement du langage et de la parole.

Une première extension possible de ce travail consisterait à enrichir la base lexicale existante par des informations complémentaires, par exemple des enregistrements sonores, ou bien encore concernant la structure morphologique des entrées lexicales. L'expérience accumulée dans le cadre de ce projet sur la représentation des entrées lexicales peut s'avérer, à cet égard, utile. Le jeu de règles produisant les variantes pourrait également être étendu pour engendrer des variantes dialectales. Concernant plus spécifiquement les applications de traitement de la parole, ce corpus gagnerait à être étendu d'une part en intégrant dans le lexique un certain nombre de groupes très fréquents dont la prononciation est extrêmement variable en fonction du débit ; d'autre part en y ajoutant des textes (textes journalistiques ou littéraires, transcriptions d'enregistrements...) phonétisés. De telles ressources textuelles sont en effet indispensables pour apprécier qualitativement les performances des systèmes de transcription en particulier en matière de détection des homographes-hétérophones, de prévision de la réalisation des liaisons, de sélection de stratégies de prononciation cohérentes, qui sont autant de difficultés rémanentes auxquelles sont confrontés les systèmes de transcription.

Ce lexique constitue une ressource dont nous pensons qu'elle sera directement utilisable pour les applications de traitement automatique de la parole. Nous espérons que son exploitation opérationnelle, que nous souhaitons la plus large possible, suscitera un retour des utilisateurs nous permettant de l'améliorer.

Remerciements

Ce travail a été financé par le programme "Ingéniérie des Langues : Production, validation et mise à disposition de données et d'outils linguistiques", thème "Valorisation de ressources et traitement de la parole".

Nous remercions F. Sannier, A. Amelot, O. Corbin et C. Rustin pour leur contribution à la production et à la vérification des transcriptions phonétiques. Merci également à F. Béchet pour la fourniture de listes de noms propres étiquetés.

Références

[ABM+95] G. Adda, P. Blache, J. Mariani, P. Paroubek, and M. Rajman. Action grace: Mise en place du paradigme d'évaluation. application au domaine de l'analyse morpho-syntaxiqe. In *TALN'95*, pages 72–79, Marseille, France, 1995.

[Aub91] V. Aubergé. *La synthèse de la parole : des règles aux lexiques.* PhD thesis, Université P. Menès-France, Grenoble, 1991.

[Bur90] G. Burnage. CELEX: a guide for users. Technical report, University of Nijmegen, Center for Lexical Information, Nijmegen, 1990.

[CM96] N. Calzolari and M. Monachini. Common specifications and notation for lexicon encoding. Technical report, Rapport MULTEXT-LRE, 1996.

[dM97] P. Boula de Mareüil. *Étude linguistique appliqueée à la synthèse de la parole à partir du texte.* PhD thesis, Université Paris XI, Orsay, 1997.

[dM98] P. Boula de Mareüil. Guide du transcripteur pour la constitution d'un lexique phonétique du français, 1998.

[GMe97] D. Gibbon, R. Moore, and R. Winski (eds). *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, Berlin, 1997.

[IV94] N. Ide and J. Veronis. MULTEXT: Multilingual text tools and corpora. In *Coling'94*, volume 1, pages 588–592, Kyoto, Japan, 1994.

[Lap89] É. Laporte. Quelques variations phonétiques en français. *Linguisticæ Investigationes*, XII(1):43–116, 1989.

[LD90] A. Lacheret-Dujour. *Contribution à une analyse de la variabilité phonologique pour le traitement automatique de la parole continue multi-locuteur.* PhD thesis, Université de Paris VII, 1990.

[PdCFP92] G. Pérennou, M. de Calmès, I. Ferrané, and J-M. Pécatte. Le projet BDLEX de base de données lexicales du français écrit et parlé. Actes du séminaire lexique, Toulouse, 1992.

[RLP97] M. Rajman, J. Lecomte, and P. Paroubek. Format de description lexicale pour le français : Description morpho-syntaxique. Technical report, Grace-GTR-3.2.1, 1997.

[SBEB+96] T. Spriet, F. Béchet, M. El-Bèze, C. de Loupy, and L. Khouri. Traitement automatique des mots inconnus. In *TALN'96*, Marseille, 1996.

[YBd+98] F. Yvon, P. Boula de Mareüil, C. d'Alessandro, V. Aubergé, M. Bagein, G. Bailly, F. Béchet, S. Foukia, J-P. Goldman, É. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, and B. Zellner. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in french. *Computer Speech and Language*, 12(4), 1998.

[Yvo96] F. Yvon. *Prononcer par analogie : motivations, formalisations et évaluations.* PhD thesis, ENST, 1996.

Production

Modélisation articulatoire linéaire 3D d'un visage pour une Tête Parlante Virtuelle

Pascal Borel, Pierre Badin, Lionel Revéret, Gérard Bailly

ICP - UMR 5009 CNRS / INPG / Université Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex 1, France
Tél.: ++33 (0)476 57 48 27 - Fax: ++33 (0)476 57 47 10
Mél: borel@icp.inpg.fr - http://www.icp.inpg.fr

ABSTRACT

This article presents 3D linear articulatory models of the face (skin and lips) for speech, based on articulatory measures extracted from video images of a French speaker. Linear statistical analysis of the 3D coordinates of lower jaw incisor, flesh-points on the skin and lip geometry has allowed to extract five degrees of freedom that account for about 96 % of the total variance of the data. Two jaw parameters correspond to jaw height and advance, while three lip / skin parameters correspond to lip protrusion, lip separation and lip height. A linear model of face controlled by these parameters has then been developed and integrated in the ICP Virtual Talking Head. The RMS reconstruction error obtained reconstructing the face with the model is about 0.1 cm.

1. INTRODUCTION

La modélisation articulatoire linéaire a été largement utilisée pour décrire le mouvement des articulateurs internes de la parole tels que la langue ou le conduit vocal ([Mer73] ; [Mae79]). Plus récemment, [Beau96] ont développé un modèle articulatoire médiosagittal de conduit vocal basé sur un film cinéradiographique tourné sur un sujet PB. Ce modèle a ensuite été généralisé à la troisième dimension en utilisant des données IRM obtenues sur le même sujet ([Bad98]).

Nous avons ici adopté une approche similaire pour développer un modèle articulatoire linéaire 3D de visage à partir de données vidéos acquises sur le sujet PB. Ainsi, les modèles des différents articulateurs de la parole (internes et externes) pourront être intégrés dans une véritable "Tête Parlante Virtuelle" et contrôlés par un même jeu de paramètres articulatoires (cf. [Bad00]). Les applications d'un tel avatar sont nombreuses : communication multimodale (labiophone), synthèse audiovisuelle à partir du texte, aide à l'apprentissage des langues, etc.

2. DONNÉES ARTICULATOIRES

La présente approche consiste à extraire les degrés de liberté des organes par analyse statistique linéaire des mesures articulatoires réalisées sur un corpus soigneusement conçu. Dans un souci de cohérence avec les données IRM de conduit vocal acquises par [Bad98], nous avons utilisé le même corpus, à savoir les 10 voyelles orales du français et les consonnes [p t k f s ʃ ʀ l] en contexte symétrique [a i u], soit un total de 34 articulations soutenues.

2.1. Acquisition des images vidéo

Le visage du locuteur a été filmé de face et de profil sous des conditions contrôlées d'éclairage. Un miroir incliné à 45° a permis d'obtenir ces deux vues sur une même image vidéo. 32 points spécifiques de la peau ont été repérés par des petites billes de plastique collées sur le visage (cf. figure 1). D'autre part, les lèvres ont été maquillées en bleu afin de bien discerner le vermillon des lèvres du reste de la peau. Enfin, une éclisse mandibulaire a été fixée à la mâchoire inférieure du sujet afin de suivre les déplacements sous-jacents de la mandibule.

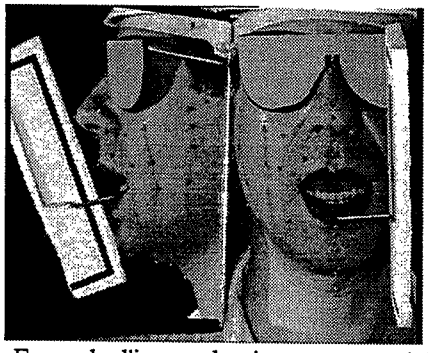


Figure 1: Exemple d'image du visage pour un /a/.

Notons également que la correspondance entre les deux vues a été calibrée grâce à un objet 3D de dimensions connues, permettant ainsi une reconstruction 3D stéréoscopique des données.

2.2. Extraction des mesures articulatoires

Le dépouillement des images acquises vise à en extraire les différentes données nécessaires au développement du modèle. Ces données sont de trois types : (1) points 3D de peau obtenus par reconstruction stéréoscopique des 32 billes collées sur le visage ; (2) points 3D de lèvres acquis en appliquant la méthode de mesure labiale présentée dans [Rev98] : un maillage ajuste globalement la forme des lèvres selon la position de 30 points de contrôle 3D ;

(3) mesures de *paramètres articulatoires globaux* de mâchoire (*JawHei*: hauteur, et *JawAdv*: avancée, mesurées sur l'incisive inférieure en utilisant l'éclisse mandibulaire), et de lèvres (*ProTop*: protrusion de la lèvre supérieure ; *LipHei*: hauteur de l'ouverture intéro-labiale ; *LipTop*: hauteur de la lèvre supérieure par rapport à l'incisive supérieure) en appliquant la méthode de [Lal91] basée sur un ChromaKey bleu.

En fait, le même corpus a été enregistré une fois avec l'éclisse mandibulaire, et une fois sans cette éclisse. A partir du corpus avec éclisse, nous avons mesuré les coordonnées 3D de l'incisive inférieure, et établi une relation de prédiction linéaire de ces coordonnées à partir de l'ensemble des points de la peau. Le corpus sans éclisse permet d'une part d'éviter les déformations des lèvres liées à l'éclisse, et d'autre part d'augmenter la résolution en cadrant le visage plus serré. Pour ce corpus, nous avons estimé la position de l'incisive inférieure par la relation établie pour le corpus avec éclisse ; nous avons vérifié que l'erreur quadratique moyenne sur le sous-ensemble des configurations pour lesquelles l'incisive est directement visible et mesurable était inférieure à 0.1 cm.

Nous disposons finalement, pour chaque configuration du corpus sans éclisse qui sera utilisé pour la modélisation, de 32+30 points 3D décrivant le visage ainsi que de 5 mesures de paramètres articulatoires globaux. Notons enfin que ces différentes données ont toutes été recalées dans un même repère absolu en appliquant, à chaque configuration, une roto-translation dont les paramètres ont été obtenus à partir de trois points supposés immobiles dont on connaît les coordonnées 3D dans le repère absolu.

3. ANALYSE STATISTIQUE ET MODÉLISATION LINÉAIRE

La modélisation articulatoire linéaire consiste à représenter des organes échantillonnés de manière fine par une combinaison d'un nombre restreint de composantes linéaires correspondant aux degrés de liberté de ces organes. Ces composantes peuvent être choisies de manière arbitraire, ou calculées par une analyse statistique linéaire telle que l'Analyse en Composantes Principales (ACP). Nous présentons dans cette section deux approches pour obtenir les paramètres de contrôle du modèle, ainsi qu'une description et une évaluation des modèles correspondants.

3.1. Première analyse des données

Dans cette première approche, deux paramètres seulement sont imposés, *jaw1* et *jaw2*, qui sont les deux premiers facteurs fournis par une ACP appliquée aux coordonnées 3D de l'incisive inférieure. Notons que ces paramètres sont relativement corrélés avec *ZJH* et *ZJA* qui sont respectivement *JawHei* centré normé, et *JA* centré normé avec *JA = JawAdv – contribution de ZJH*. La mâchoire étant l'organe qui porte la lèvre inférieure, *jaw1* est imposé comme premier prédicteur. Par ailleurs, *JawAdv* étant en partie corrélé avec les différentes mesures labiales globales, *jaw2* est imposé comme prédicteur des résidus obtenus après soustraction des contributions de trois paramètres labiaux (*lip1*, *lip2*, *lip3*) obtenus par ACP des 30 points de lèvres. Enfin, un paramètre de peau supplémentaire (*skin1*) est obtenu par ACP sur les points de peau, après avoir retiré les contributions de *jaw1*, *lip1*, *lip2*, *lip3* et *jaw2* dans cet ordre. Par la suite, ces paramètres de commande du modèle de visage seront appelés *paramètres AudioVidéo (AV)*.

3.2. Analyse guidée par les mesures de paramètres articulatoires globaux

Pour cette seconde analyse, nous avons imposé comme prédicteurs les mesures labiales globales à la place des paramètres *lip1*, *lip2* et *lip3* obtenus par ACP. Ces nouveaux paramètres de commande du modèle, notés *ZLP*, *ZLH* et *ZLV*, sont respectivement *LP*, *LH* et *LV* centrés normés avec :
LP = ProTop – contribution de ZJH ;
LH = LipHei – contribution de ZJH ;
LV = LipTop – contrib. de ZJH – contrib. de ZLP et ZLH.
Par ailleurs, les paramètres *ZJH* et *ZJA* remplacent respectivement *jaw1* et *jaw2*, tandis qu'un paramètre supplémentaire *ZSK* est déterminé de manière similaire à *skin1*. Par la suite, les paramètres de commande *ZJH*, *ZLP*, *ZLH*, *ZLV*, *ZJA* et *ZSK* seront appelés *paramètres MédioSagittaux (MS)*.

3.3. Modèles et évaluation

Deux modèles de visage ont ainsi été développés, correspondant à chacune des analyses présentées précédemment. Ces modèles linéaires sont entièrement définis par la moyenne de chacune des coordonnées 3D des points du visage, et par la matrice des coefficients qui prédisent l'écart de ces points par rapport à leur moyenne comme combinaison linéaire des six paramètres de commande considérés. Dans le cas où ces paramètres sont non corrélés, les coefficients de prédiction sont calculés par régression linéaire multiple sur l'ensemble du corpus entre les données centrées et les paramètres de commande. Les paramètres de notre étude étant partiellement corrélés, la régression multiple a été décomposée en une succession de régressions linéaires simples entre chacun des paramètres et le résidu des données centrées obtenu après soustraction des contributions des paramètres précédents.

Afin d'étudier les effets de chacun des paramètres de commande, nous avons réalisé des nomogrammes pour chaque paramètre, en faisant varier ce paramètre entre les valeurs normalisées –3 et +3, tout en maintenant les autres paramètres à zéro (voir les nomogrammes pour les paramètres MS à la figure 2). Nous avons constaté que les nomogrammes des deux modèles sont extrêmement similaires. En particulier, les paramètres *lip1*, *lip2*, *lip3* sont assez fortement corrélés avec *ZLP*, *ZLH*, *ZLV* respectivement (coefficients de corrélation de 0.98, 0.89 et 0.86). Il est intéressant de remarquer que ces

paramètres, qui représentent les degrés de liberté du système articulatoire lèvres / peau, correspondent parfaitement aux différents traits phonétiques traditionnellement utilisés pour la labialité (cf. [Abr86]). Le premier paramètre labial *ZLP / lip1* correspond clairement à un effet de protrusion – arrondissement ; le deuxième paramètre *ZLH / lip2* correspond à un mouvement d'aperture ; le troisième paramètre *ZLV / lip3* correspond à un mouvement vertical quasi simultané des deux lèvres qui permet en particulier la réalisation de lèvres avancées et ouvertes pour les consonnes /ʃ ʒ/ ou pour les labio-dentales. Notons enfin que le paramètre *skin1 / ZSK* contrôle en partie la position de la pomme d'Adam, qui marque la position de la boîte laryngée.

Les pourcentages de la variance globale des données de visage expliquée par chacun des paramètres de commande pour les deux analyses précédentes sont reportés dans la table 1. Ces résultats montrent à nouveau une similitude entre les deux types d'analyses. Cependant, le pourcentage total de la variance expliquée par ACP est légèrement plus important que celui expliqué par l'analyse guidée, ce qui n'est pas surprenant puisque l'ACP est optimale en terme d'explication de variance. La figure 3 illustre de manière plus détaillée la réduction de variance induite par la soustraction des contributions des six paramètres aux mesures originales.

Table 1: Pourcentages de la variance globale des données de visage expliquée par chacun des paramètres de commande du modèle.

Paramètres de commande	Paramètres AV Variance (%)	Paramètres MS Variance (%)
jaw1 / ZJH	16.20	16.36
lip1 / ZLP	74.42	72.08
lip2 / ZLH	3.74	3.03
lip3 / ZLV	2.18	1.67
jaw2 / ZJA	0.30	1.02
skin1 / ZSK	0.82	1.64
Total	97.66	95.80

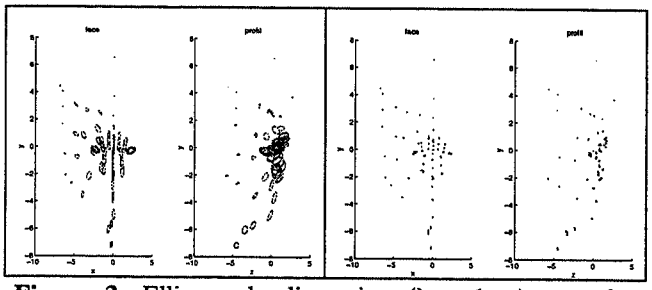


Figure 3: Ellipses de dispersion (à ± 1 σ) pour les mesures originales (gauche) et pour leurs résidus après soustraction des contributions des six paramètres (droite).

Enfin, une évaluation de la précision des deux modèles de visage a été réalisée en déterminant, pour chacun des modèles, l'erreur quadratique moyenne totale de prédiction des points du visage à partir de 5 paramètres de commande. La reconstruction du visage à partir des paramètres AV présente une RMS de 0.07 cm. En utilisant les paramètres MS comme commandes du modèle linéaire, la RMS obtenue est alors de 0.09 cm. Outre l'erreur de reconstruction plus faible, le premier modèle présente l'avantage de ne pas nécessiter de mesure explicite des paramètres labiaux globaux. Néanmoins, les deux types de paramètres de contrôle sont acceptables.

4. Relations entre modèles de visage et modèles de conduit vocal

Dans le cadre du projet "Tête Parlante Virtuelle", l'ICP développe des modèles articulatoires linéaires des différents articulateurs de la parole commandés par un même jeu de paramètres (cf. [Bad00]), les paramètres MS, qui constituent un sous-ensemble des paramètres de contrôle du modèle articulatoire médiosagittal développé par [Beau96] à partir de données cinéradiographiques du sujet PB. Dans le cadre de ce projet, il est intéressant de pouvoir contrôler le modèle de visage à partir des paramètres MS, mais nous avons constaté que l'erreur de reconstruction des données était légèrement plus faible pour le contrôle à partir des paramètres AV. Nous avons donc comparé les erreurs de reconstruction des points de visage suivant deux méthodes : (1) prédiction directe à partir des paramètres MS (RMS de 0.09 cm), et (2) prédiction à partir des paramètres AV déduits par transformation linéaire des paramètres MS (RMS de 0.11 cm). Les deux méthodes conduisent à des erreurs relativement comparables.

5. Conclusions et perspectives

Dans cet article, nous avons présenté un modèle articulatoire linéaire 3D de visage (lèvres + peau) contrôlé par cinq paramètres articulatoires. La seule autre approche similaire, à notre connaissance, est celle de [Yeh98] pour un locuteur anglais et pour un locuteur japonais ; ils n'ont cependant pas cherché à obtenir des paramètres de contrôle clairement interprétables en termes articulatoires.

Ce type de modèle développé à partir de données articulatoires précises permet une grande qualité de synthèse visuelle comme en témoignent les résultats obtenus par [Rev00], qui ont habillé le présent modèle articulatoire par plaquage et mélange de textures extraites d'images du même sujet, et obtenu un visage synthétique beaucoup plus réaliste que ceux obtenus par d'autres techniques. Les autres modèles classiques en synthèse audiovisuelle tels que les modèles topologiques de visage ou ceux basés sur le "morphing" entre images cibles correspondant à des visèmes (cf. [Ben98]) présentent cependant l'avantage de pouvoir être facilement adaptables à des physionomies différentes.

Rappelons que notre modèle de visage est contrôlé par des paramètres articulatoires pouvant être reliés simplement aux paramètres de commande d'autres modèles articulatoires linéaires développés sur le même sujet, ce qui nous permet une intégration dans la "Tête Parlante Virtuelle" en cours de développement à l'ICP ([Bad00]).

Enfin, ce modèle articulatoire de visage constitue la base de projets de visiophonie (transmission de visages à bas débit) dans lesquels l'ICP est fortement impliqué. Dans ce cadre, l'une des extensions du présent travail consistera à développer des modèles sur d'autres sujets afin de constituer une base permettant de s'adapter à n'importe quel autre locuteur. Enfin, ce type d'approche permettra également d'intégrer, en suivant la même méthodologie, des expressions telles que le sourire dans la modélisation du visage.

REMERCIEMENTS

Ce travail a été mené dans le cadre du programme de recherche "Une Tête Parlante Virtuelle : Données et modèles en production de parole" financé par l'Agence Rhône-Alpes pour les Sciences Sociales et Humaines. Nous sommes reconnaissants pour leur aide à nos collègues de l'ICP (en particulier A. Arnal et C. Savariaux), et au docteur G. Rozencweig pour la réalisation de l'éclisse mandibulaire.

RÉFÉRENCES

[Abr86] Abry C., Boë L.J. (1986), "'Laws' for Lips", Speech Communication, Vol. 5, pp. 97-104.

[Bad98] Badin P., Pouchoy L., Bailly G., Raybaudi M., Segebarth C., Lebas J.F., Tiede M., Vatikiotis-Bateson E., Tohkura Y. (1998), "Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM", 22èmes JEP, Martigny, Suisse, pp. 283-286.

[Bad00] Badin P., Borel P., Bailly G., Revéret L. (2000), "Towards an Audio-visual Virtual Talking Head: 3D linear articulatory modelling of tongue, lips and face based on MRI and video images", 5th Speech Production Seminar, Munich, Allemagne (accepté).

[Beau96] Beautemps D., Badin P., Bailly G., Galvan A., Laboissière R. (1996), "Evaluation of an articulatory-acoustic model based on a reference subject", 4th Speech Production Seminar / ETRW, pp. 45-48.

[Ben98] Benoît C., Le Goff B. (1998), "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP", Speech Communication, Vol. 26, pp. 117-129.

[Lal91] Lallouache M.T. (1991), "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", Thèse de doctorat, INPG, Grenoble, France.

[Mae79] Maeda S. (1979), "Un modèle articulatoire de la langue avec des composantes linéaires", 10èmes JEP, Grenoble, France, pp. 152-162.

[Mer73] Mermelstein P. (1973), "An articulatory model for the study of speech production", J. Acoust. Soc. Am., Vol. 53, pp. 1070-1082.

[Rev98] Revéret L., Benoît C. (1998), "A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production", AVSP'98, Terrigal, Australie, pp. 207-212.

[Rev00] Revéret L., Bailly G., Borel P., Badin P. (2000), "Analyse par la synthèse d'un visage 3D parlant : inversion optico-articulatoire", 23èmes JEP, Aussois, France (accepté).

[Yeh98] Yehia H., Rubin P., Vatikiotis-Bateson E. (1998), "Quantitative association of vocal-tract and facial behavior", Speech Communication, Vol. 26, pp. 23-43.

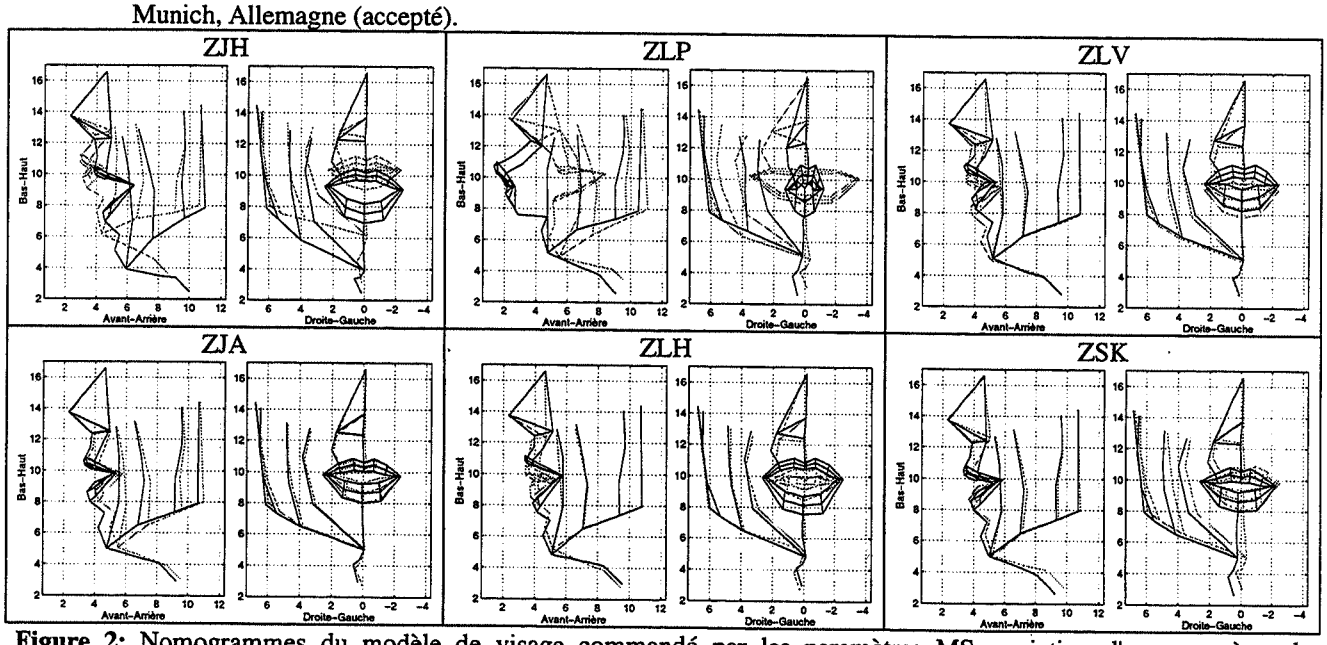


Figure 2: Nomogrammes du modèle de visage commandé par les paramètres MS: variation d'un paramètre de commande entre les valeurs –3 (pointillés) et +3 (continu), les autres paramètres étant maintenus à zéro.

Analyse par la synthèse d'un visage 3D parlant : inversion optico-articulatoire

Lionel Revéret, Gérard Bailly, Pascal Borel, Pierre Badin

ICP – CNRS UMR 5009/ Université Stendhal / INPG
46, av Félix Viallet, 38031 Grenoble CEDEX 01, France
Tél.: ++33 (0)476 57 45 40 - Fax: ++33 (0)476 57 47 10
Mél: reveret@icp.inpg.fr - http://www.icp.inpg.fr/~reveret

ABSTRACT

We present a method for the automatic analysis and 3D synthesis of a video sequence of a French speaker. The synthesis is based on an articulatory model of the face, controlled by 6 parameters. A texture mapping algorithm achieves a video realistic rendering by an "alpha blending" technique. This video-realistic model of talking face is used for automatic analysis of a video sequence of a speaker. A comparison is performed directly at the pixel level between the original image and the synthesis result in order to estimate the 6 control parameters and the head orientation. This process implements an articulatory inversion of the face model directly from the image signal. A quantitative evaluation of this process is described.

1. INTRODUCTION

Pour des applications de télécommunications (téléconférence, animation de clone virtuel) comme pour l'analyse fine de corpus audiovisuels, le besoin en modèle paramétré est crucial pour l'analyse automatique de visages parlants. En effet, surtout en l'absence de maquillage ou de marqueurs, l'analyse ascendante (des pixels vers la forme) des images d'un locuteur devient très imprécise : faiblesse des contrastes, variabilité des conditions. Elle nécessite alors une régularisation qui peut être résolue par projection/inversion de modèle sur l'image (analyse descendante) afin d'extraire des paramètres de forme [Rev99]. Au cours des deux dernières décennies, l'animation faciale paramétrée a été marquée par trois grandes approches :

- l'approche géométrique où le contrôle est assuré par des paramètres mesurables, directement liés à un modèle topologique du visage [Par91; Ben98],
- l'approche physiologique contrôlée par des modèles de simulation de l'action des muscles faciaux et de l'élasticité de la peau [Ter90],
- l'approche par déformation d'images (« morphing ») où des images réelles de visages sont modifiées au niveau du pixel [Ezz98].

Le gain en réalisme qu'apporte l'approche physiologique par rapport à l'approche géométrique se fait au détriment d'une plus grande complexité dans l'organisation du contrôle. De plus, la difficulté pour modéliser des muscles sans insertion osseuse, tel que *l'orbiculoris oris*, rend délicate l'utilisation de ces modèles pour la synthèse visuelle de la parole.

L'approche par « morphing » offre des possibilités de réalisme vidéo intéressantes puisqu'elle se base sur des images réelles de visages. Elle s'affranchit le plus souvent de modèles paramétrés du visage : seule est contrôlée la transition continue entre des images cibles correspondant à des visèmes. Cette approche limite la synthèse aux conditions de la session enregistrée et nécessite de définir la transition entre chaque paire de visèmes.

Les modèles statistiques linéaires ont largement été utilisés pour la description de la géométrie interne du conduit vocal et son contrôle selon peu de paramètres articulatoires. Nous proposons ici une approche similaire de description articulatoire pour un modèle 3D de visage parlant. Ce modèle est couplé avec une méthode de synthèse vidéo réaliste par plaquage de textures, inspirée des principes de synthèse par « morphing ». La minimisation de la différence entre image originale du locuteur et synthèse réalise alors *une analyse automatique de visage parlant totalement descendante*, où le signal image est utilisé directement au niveau pixel pour inverser les paramètres du modèle articulatoire. Une évaluation quantitative de cette méthode est proposée, basée sur la différence d'images, pour un locuteur maquillé en bleu. La même méthode est applicable à un locuteur non maquillé et sera évaluée dans des travaux futurs.

2. DONNEES D'APPRENTISSAGE DU MODELE

Les données d'apprentissage nécessaires à la définition du modèle articulatoire par analyse statistique sont issues du projet « Tête parlante » de l'ICP [Bad00]. Nous donnons ici un rapide descriptif des données pour le visage.

Le visage du locuteur a été filmé de face et de profil sous des conditions contrôlées d'éclairage (lampe de 1000W). Un miroir incliné à 45° a permis d'obtenir sur la même image vidéo en synchronie les vues de face et de profil. 34 points de contrôle ont été repérés par des billes de plastiques collées sur la partie droite du visage. Afin de permettre une reconstruction 3D stéréoscopique, la correspondance entre les deux vues a été calibrée grâce à un objet de dimensions connues. Les lèvres ont été peintes en bleu afin de réduire l'ambiguïté dans la détermination des contours. La méthode de mesure des lèvres présentée dans [Rev98] a été utilisée pour obtenir 30 points 3D supplémentaires : un maillage ajuste la forme des lèvres selon la positon des 30 points de contrôle. Ainsi pour une

image vidéo, tout le visage du locuteur a été mesuré selon 64 points 3D, soit 192 coordonnées XYZ. Le corpus est constitué d'un ensemble de 34 images correspondants au maximum de réalisation de :

- 10 voyelles, V∈ {a, i, y, u, o, e, ɔ, ø, ɛ, œ},
- 8 consonnes dans 3 contextes vocaliques, vCv avec C∈ {p, ʃ, ʀ, l, f, k, t, s}, v∈ {a, i, y}.

3. MODELE ARTICULATOIRE 3D

Dans la perspective de l'appliquer à l'analyse automatique et ainsi conserver le maximum de variance des données, les paramètres de contrôle sont issus directement de trois analyses en composantes principales (ACP) « guidées », i.e. visant à séparer itérativement l'influence de la mâchoire, des lèvres et des muscles faciaux (voir Borel et al., dans ce volume). Le nombre de paramètres retenus dans chaque ACP est respectivement 2, 3 et 1 (table 1).

Les figures 1 à 3 présentent les nomogrammes des 6 paramètres. Certains points du visage ont été reliés entre eux afin de clarifier la lecture. Chaque figure de face et profil correspond à une variation de ± 3 fois l'écart-type.

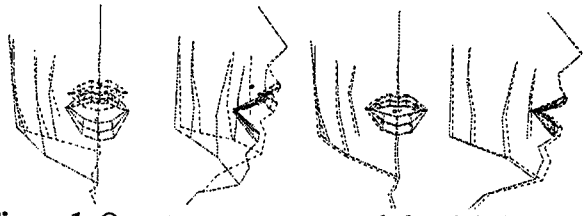


Figure 1: Ouverture et avancement de la mâchoire.

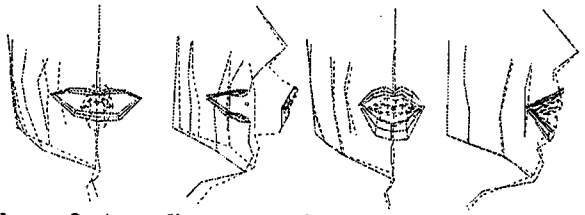


Figure 2: Arrondissement et fermeture des lèvres.

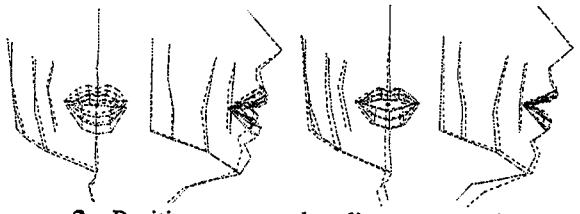


Figure 3: Positionnement des lèvres pour les labio-dentales ; Abaissement de la pomme d'Adam.

Table 1: Part de la variance des 34 formes expliquée par chacun des 6 paramètres du modèle

	Variance (%)	Somme cumulée
1. mâchoire (1)	18.0	18.0
2. mâchoire (2)	0.4	18.4
3. lèvres (1)	72.6	91.0
4. lèvres (2)	3.8	94.8
5. lèvres (3)	2.1	96.9
6. face (1)	0.8	**97.7**

Les noms donnés a posteriori aux composantes ne visent qu'à fournir des interprétations qualitatives des gestes. Plutôt qu'une description biomécanique exacte des articulateurs, ils expriment la cinématique résultante des couplages fonctionnels entre ces articulateurs. Ces résultats montrent la prépondérance du geste d'arrondissement / protrusion des lèvres. Nous avons retenu le dernier paramètre pour son interprétation articulatoire, bien qu'il n'explique qu'une part faible de la variance totale des données.

4. SYNTHESE PAR PLAQUAGE DE TEXTURE

Le modèle articulatoire génère la position de 64 points 3D à partir de 6 paramètres. Un maillage polygonal, construit sur ces derniers, a été développé afin de fournir un rendu visuel de toute la surface du visage. Par une technique de plaquage de texture, une image réelle du locuteur a été appliquée sur le maillage.

4.1. Définition d'un maillage et « morphing »

Le maillage des lèvres est issu d'une interpolation polynomiale des 30 points de contrôle [Rev98]. La densité réglable a été fixée à 144 polygones quadrilatères. Pour le reste du visage, aucun point n'a été ajouté. Un maillage de 39 polygones triangulaires a été défini. Les polygones joignant les lèvres au reste du maillage ont été affinés afin d'assurer la continuité de la surface.

Grâce à la bibliothèque de développement d'applications graphiques OpenGL, une fois établie la correspondance entre les points du maillage et leur position sur l'image, les déformations de l'image suivent automatiquement celles du maillage par interpolation des pixels (figure 4). Cette technique de synthèse permet de retrouver une image d'apparence réaliste malgré un maillage de faible densité. De plus, des cartes graphiques accélératrices 3D standard assurent une synthèse en temps réel sur PC de cette technique.

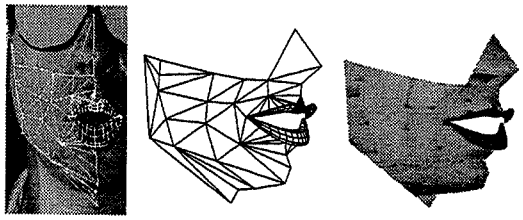


Figure 4: Correspondance maillage/texture ; maillage 3D cible ; résultat du plaquage de texture sur le maillage.

4.2. Synthèse par mélange de textures

Malgré le plaquage de texture, certains détails de la surface du visage ne peuvent pas être correctement rendus en raison de la faible densité du maillage. Typiquement, les effets d'ombre dus au pli naso-génien (entre joue et bouche) ne sont pas rendus. Ce pli est particulièrement saillant lors de la réalisation de voyelles étirées. Or, sur la texture d'origine choisie où la voyelle est arrondie, ce pli n'apparaît pas. Aussi, malgré la forme étirée du maillage,

ce pli n'est pas visible après « morphing », étant absent de la texture d'origine (figure 5.a).

Pour résoudre ce problème, au lieu d'une seule texture, 5 formes de référence les plus éloignées ont été utilisées et mélangées à la synthèse par combinaison linéaire (« alpha blending »). Pour chacune, le maillage a été mis en correspondance avec l'image correspondante.

Soient $M_{i=1..5}$ les 5 maillages 3D et $T_{i=1..5}$ les 5 textures. Soit S la fonction graphique de plaquage de texture qui à tout maillage 3D M et à tout couple $[M_i ; T_i]$ associe la synthèse d'une image. L'image finale I est égale à :

$$I = \sum_{i=1}^{5} \alpha_i S(M, [M_i ; T_i]), \text{ avec } \alpha_i = e^{-k_i d(M, M_i)}$$

d désigne la distance euclidienne entre les points du maillage. Les coefficients k_i sont obtenus par optimisation de la reconstruction des maillages $M_{j=1..34}$ des 34 visèmes par une combinaison linéaire similaire :

$$\hat{M}_{j=1..34} = \sum_{i=1}^{5} e^{-k_i d(M_j, M_i)} M_i \text{ et}$$

$$k_i = \arg\min\left(\left\| M_j - \hat{M}_j(k) \right\|^2_{j=1..34} \right)$$

La figure suivante montre la correction apportée par cette méthode, notamment pour le pli naso-génien (fig. 5.b).

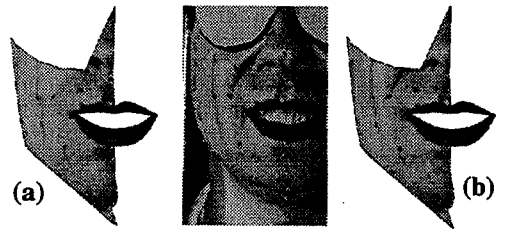


Figure 5: Plaquage d'une seule texture ; image originale ; plaquage et mélanges pondéré de 5 textures.

5. INVERSION AUTOMATIQUE DU MODELE A PARTIR DE LA VIDEO

L'analyse automatique d'une séquence vidéo du locuteur est effectuée en comparant directement, pixel à pixel, chaque image originale de la séquence avec le résultat de synthèse par mélange de textures (figure 6). Une inversion du modèle contrôlé par 6 paramètres est ainsi opérée à partir du signal de l'image. Les 6 paramètres de position de la tête (3 rotations, 3 translations) sont aussi ajustés automatiquement par cette procédure.

Figure 6: Convergence du modèle sur l'image.

Le critère d'erreur entre l'image originale et le résultat de synthèse du modèle est égal à la moyenne sur l'image des différences en valeur absolue des niveaux RGB. Un algorithme de gradient conjugué optimise la valeur des 12 paramètres de contrôle pour minimiser cette erreur.

6. RESULTATS, EVALUATIONS ET DISCUSSIONS

6.1. Influence de la synthèse par mélange de textures

La figure 7 compare la synthèse du modèle par plaquage d'une texture unique et la synthèse par mélange de textures. Pour les 34 visèmes d'apprentissage, le maillage provient d'un étiquetage manuel des 64 points 3D. La moyenne des différences sur les 34 visèmes est présentée. Pour chaque visème, le résultat de la différence entre image originale et image synthétisée a été déformé à nouveau vers la forme moyenne, ceci afin de permettre un alignement cohérent des pixels entre les 34 formes.

Figure 7: Texture unique *vs* mélange de textures.

Un pixel foncé correspond à une grande différence. La dynamique des données correspond à une quantification en 255 niveaux. Par souci de lisibilité, la dynamique des images a été modifiée : la dynamique de 0 à 255 sur l'image présentée correspond à un écart réel de 0 à 51 pour le calcul des différences. Les images apparaissent donc ici plus sombres qu'elles ne le sont en réalité. On constate une amélioration générale : différence moyenne de l'image égale 10.0±1.1 pour le mélange contre 13.0±1.9 pour une texture unique. Localement, le pli naso-génien est nettement amélioré. Restent des erreurs importantes dans la zone du nez, provenant du fait que le maillage est trop large à cet endroit.

6.2. Influence du modèle articulatoire

Le modèle articulatoire explique 97.7% de la variance totale des visèmes. La figure 8 compare la synthèse par mélange de textures entre un maillage provenant de l'étiquetage manuel de la position des 64 points 3D et leur prédiction par les 6 paramètres articulatoires.

Figure 8: Positions 3D originales *vs* prédites par modèle.

La différence moyenne passe de 10.0±1.1 pour l'étiquetage orignal à 11.4±0.9 pour la prédiction par modèle. Localement, la zone des billes apparaît. Leur

couleur étant marquée par rapport à la peau, l'erreur sur la position 3D prédite par les 6 paramètres entraîne une erreur visible sur la texture.

6.3. Analyse automatique : /apa ipi upu/

L'analyse automatique d'une séquence /apa ipi upu/ contenant 169 images a été évaluée (figures 9 et 10 ci-contre). De la même manière, la moyenne des différences a été calculée. La valeur moyenne des différences sur les 169 images est de 12.8±0.8.

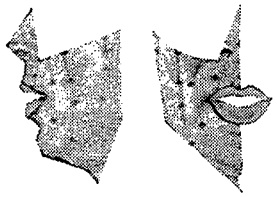


Figure 9: Différence moyenne des résultats de l'analyse automatique d'une séquence.

7. CONCLUSIONS

Nous avons présenté dans cet article trois résultats :

- un modèle 3D de visage parlant contrôlé par 6 paramètres articulatoires,
- un habillage vidéo réaliste par une méthode de synthèse par plaquage et mélange de textures,
- une analyse automatique par inversion de ce modèle texturé à partir d'une séquence vidéo.

Nos résultats visent à montrer qu'une modélisation attentive des articulateurs peut permettre de générer une synthèse réaliste de la parole visuelle à partir de peu de paramètres. De plus, la synthèse par « morphing » permet d'aborder une analyse automatique par inversion de modèle directement à partir du signal image. Ces résultats ont été obtenus pour un locuteur maquillé et muni de marqueurs. Bien que la méthode d'analyse n'en fait pas une utilisation explicite, leur présence favorise indirectement les résultats. Il reste donc à appliquer cette approche sur des séquences sans maquillage ni marqueurs.

Ce travail a été mené dans le cadre du projet « Etude d'un modèle de lèvres parlantes » financé par le CNET (réf. 991B508) et du projet « Une Tête Parlante Virtuelle » de l'Agence Rhône-Alpes pour les Sciences Sociales et Humaines.

La figure 10 présente les résultats du suivi automatique au cours du temps.

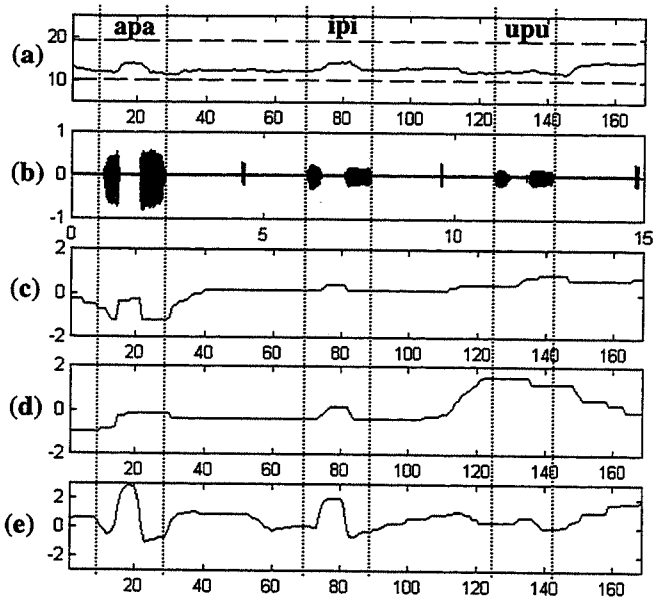


Figure 10: Résultats du suivi automatique.

(a) : Moyenne sur l'image de la différence image réelle/synthèse, la ligne pointillée inférieure montre la différence optimale de 10.0 (moyenne des visèmes), la ligne supérieure une valeur de 19.4 (moyenne sur la séquence de la différence entre l'image réelle et la forme moyenne du modèle, i.e. l'inversion des 6 paramètres articulatoires n'est pas calculée).
(b) : Signal acoustique.
(c) : Paramètre d'ouverture de la mâchoire.
(d) : Paramètre d'arrondissement des lèvres.
(e) : Paramètre de fermeture des lèvres.

Ce résultat montre une augmentation de l'erreur pour les plosives dans les cas /apa/ et /ipi/. L'état du modèle actuel ne gère aucune collision entre les lèvres supérieurs et inférieures. Ces résultats peuvent être attribués à ce manque de réalisme et seront étudiés par la suite. La mâchoire s'ouvre pour la voyelle /a/. La protrusion des lèvres apparaît très tôt pour la voyelle /u/. Les lèvres se ferment nettement sur les occlusives /p/.

BIBLIOGRAPHIE

[Bad00] Badin P., Borel P., Bailly G., Revéret L. (2000) " Towards an Audio-visual Virtual Talking Head: 3D linear articulatory modelling of tongue, lips and face based on MRI and video images ", 5th Speech Production Seminar, Munich.

[Ben98] Benoît C., Le Goff B. (1998) " Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP ", Speech Communication Journal, 26:117-129.

[Ezz98] Ezzat T., Poggio T. (1998) "MikeTalk: A Talking Facial Display Based on Morphing Visemes", Computer Animation Conference, Philadelphia.

[Par91] Parke F.I. (1991), "Control parametrization for facial animation", *in* Computer Animation'91, Springer-Verlag.

[Rev98] Revéret L., Benoît C. (1998), " A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production ", ESCA - AVSP'98.

[Rev99] Revéret L., (1999), " Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole ", Thèse de doctorat, INPG.

[Ter90] D. Terzopoulos, K. Waters, (1990) "Physically-based facial modeling, analysis, and animation", J. of Visualization and Computer Animation, 1:73-80.

Contribution à l'analyse acoustique du conduit vocal

X. Pelorson [1], *K. Motoki* [2], *R. Laboissière* [1]

[1] Institut de la Communication Parlée
CNRS, INPG, Université Stendhal, 46 avenue F. Viallet F-38031, Grenoble Cedex 01

[2] Faculty of Engineering
Hokkai-Gakuen University, S-26, W-11, Sapporo, Japan

ABSTRACT

In this paper an analysis of the vocal tract acoustics at high frequencies, including higher acoustical propagating modes, is presented.

Several examples using vocal tract approximations of increasing complexity are presented and discussed. The effects of the source position as well as of the vocal tract geometry are shown to be of particular importance.

1. INTRODUCTION

La plupart des modèles physiques appliqués au conduit vocal reposent sur une description de la propagation des ondes acoustiques en ondes planes, c'est-à-dire unidimensionnelle (e.g. [Fan60]). La limite de validité de l'hypothèse unidimensionnelle, bien que formellement difficile à déterminer (elle dépend en grande partie de la géométrie précise du conduit vocal), peut être estimée aux alentours de 4-5 kHz. Cette limitation, longtemps justifiée par le manque de données anatomiques en trois dimensions, devrait cependant pouvoir être levée grâce aux techniques modernes d'imagerie (IRM etc...). Dans cet objectif, nous présentons donc une analyse de la propagation acoustique en trois dimensions qui ne se limite donc plus aux basses fréquences. Il est montré que les effets principaux aux hautes fréquences sont liés à l'apparition de modes d'ordre supérieur qui peuvent avoir une influence considérable sur le champ de pression acoustique. Ces effets sont illustrés et discutés sur la base de quelques exemples synthétiques.

2. ELÉMENTS THÉORIQUES, PROPAGATION DANS UN CONDUIT UNIFORME

A titre d'illustration, nous considérons par la suite la représentation la plus schématique possible du conduit vocal : un guide de section rectangulaire et uniforme, de longueur L_z = 16.8 cm, rayonnant dans un écran infini. Les parois du conduit vocal sont supposées parfaitement réfléchissantes et l'excitation est assimilée à une source unique localisée en un point du conduit vocal: $Q = Q_0.\delta(z-z_0).\delta(y-y_0)\delta(x-x_0)$. L'extension de cette analyse aux cas impliquant la présence de plusieurs sources de son dans le conduit vocal ou de sources distribuées dans l'espace est aisée du fait de la linéarité du problème.

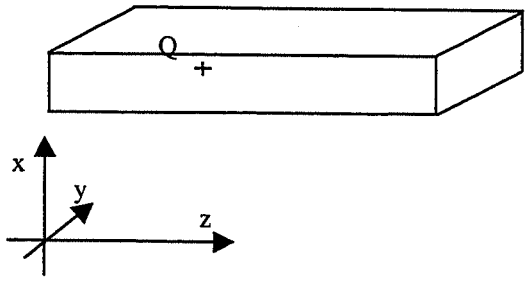


Figure 1: Représentation schématique du conduit vocal comme un guide uniforme.

La solution générale de l'équation de propagation acoustique peut alors être formulée de la manière suivante [Bru98]:

$$P(f,x,y,z)=\sum_{m,n}\Lambda_{mn}Q\frac{\Psi_{mn}(x,y)\Psi^{*}_{mn}(x_0,y_0)}{k_{mn}^2-k^2}\times \left(A_{mn}e^{-jk_{mnz}z}+B_{mn}e^{jk_{mnz}z}\right) \quad (1)$$

où f est la fréquence imposée par la source Q, A_{mn} et B_{mn} deux constantes déterminées par les conditions aux limites à chaque extrémité du guide, Λ_{mn} une constante de normalisation. Les fonctions propres Ψ_{mn} sont définies par:

$$\Psi_{mn}(x,y)=\cos\left(\frac{m\pi}{L_x}x\right)\cos\left(\frac{n\pi}{L_y}y\right) \quad (2)$$

La constante de propagation k_{mn} est déterminée par l'équation de dispersion:

$$k_{mn}^2=k^2-\left(\frac{m\pi}{L_x}\right)^2-\left(\frac{n\pi}{L_y}\right)^2 \quad (3)$$

Les équations (1) et (3) montrent qu'un mode (m,n) ne sera propagatif (k_{mn} réel) que si la fréquence d'excitation, f = kc/2π, est supérieure à la fréquence de coupure du mode, f_{mn}:

$$f_{mn}=\frac{c}{2\pi}\sqrt{\left(\frac{m\pi}{L_x}\right)^2+\left(\frac{n\pi}{L_y}\right)^2},$$

où c la vitesse du son.

En dessous de cette fréquence de coupure, k_{mn} devient imaginaire pur traduisant une onde évanescente. En particulier, lorsque la fréquence de l'excitation est inférieure à la première fréquence de coupure (f_{01} ou f_{10}) seul le mode plan (0,0) se propage.

Bien que l'équation (1) implique une double sommation infinie sur m et n, dans la pratique il n'est pas nécessaire de prendre un compte un grand nombre de modes d'ordre supérieur compte tenu de la rapidité de l'atténuation des modes évanescents. A titre d'exemple, il est présenté sur la figure 2 les contours de pression calculés à l'intérieur du guide en prenant en compte deux modes seulement. Pour comparaison, sur la même figure, sont représentés les résultats obtenus par simulation numérique (méthode TLM, [El M98]) pour la même configuration. Compte tenu de la résolution spatiale utilisée, on peut estimer que le nombre de modes pris en compte par la simulation numérique TLM est de 400.

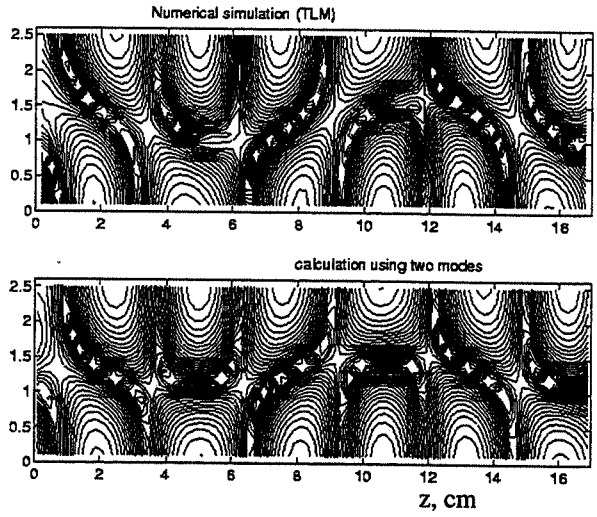


Figure 2 : Contour de pression (par pas d'1 dB) à 9 kHz obtenu par simulation numérique (en haut) et par le calcul en tenant compte de deux modes (bas).

Il est important de noter l'influence de la position de la source sur la génération et l'efficacité des modes d'ordre supérieurs. Ceci s'illustre dans l'équation (1) par le terme $\Psi^*_{mn}(x_0, y_0)$. Ainsi si la source est placée au centre d'une section du conduit vocal ($x_0 = L_x/2$ ou $y_0 = L_y/2$), comme pour la glotte en première approximation, les modes impairs ne seront pas générés ($\Psi^*_{mn}(x_0, y_0) = 0$). Réciproquement une source placée près d'une paroi excitera les modes impairs avec une amplitude maximum ($|\Psi^*_{mn}(x_0, y_0)| = 1$).

3. CHANGEMENT DE SECTION

Afin d'illustrer la génération et la transmission des modes d'ordre supérieur, nous considérons maintenant l'approximation décrite dans la figure 2 (approximation à deux sections du conduit vocal). Deux cas sont considérés, l'un parfaitement symétrique et l'autre asymétrique.

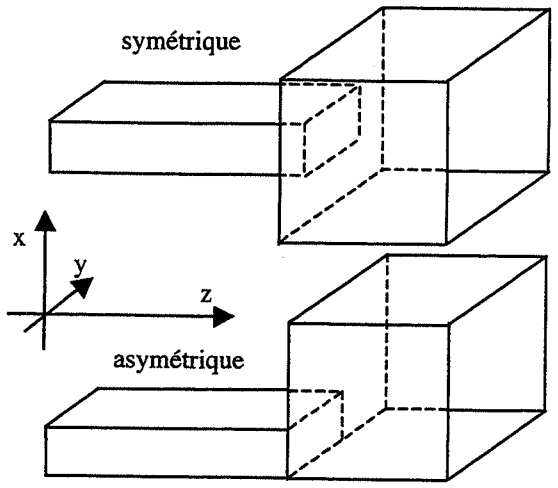


Figure 3: Approximations à deux tubes du conduit vocal. Configurations symétrique (en haut) et asymétrique (en bas).

La résolution de l'équation de propagation acoustique dans de telles structures n'est, en général, pas analytique. La méthode de raccordement modal [Ker91] permet cependant d'obtenir numériquement une solution. Le principe consiste à développer pression et vitesse acoustique sur la base des fonctions propres de chacune des deux sections, l'application du principe de la conservation de la pression et de la vitesse normale à la jonction des deux guides permet alors de déterminer l'amplitude de la pression en tout point de la structure. Dans la figure 4 sont présentés deux exemples de résultats pour les distribution de pression dans le plan z-y à 6 kHz.

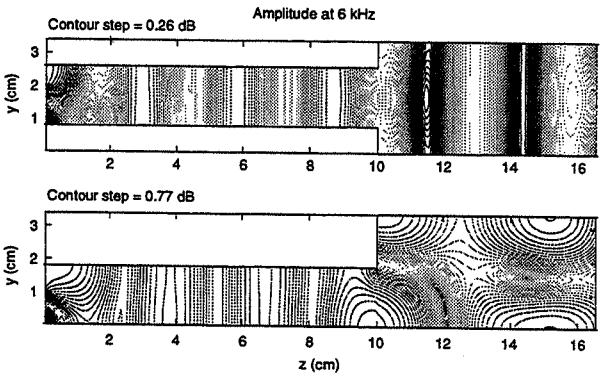


Figure 4 : Comparaison entre les contours de pression obtenus à 6 kHz dans les configuration symétriques et asymétriques.

Dans le cas d'une configuration symétrique, on peut observer que si des perturbations liées à la génération de modes d'ordre supérieurs interviennent au niveau de la source ainsi qu'aux discontinuités de sections, elles décroissent rapidement avec la distance et la propagation devient quasi-unidimensionnelle.

Dans le cas d'une configuration asymétrique on observe la génération et la propagation du mode (0,1) dans la seconde section. Ce résultat qui peut sembler surprenant, de prime abord, peut être expliqué théoriquement de

manière analogue à celle de l'influence de la position de la source.

4. APPROXIMATIONS PLUS RÉALISTES

Nous abordons enfin les calculs effectués pour deux configurations plus réalistes. Dans les deux cas les données géométriques ont été estimées à partir d'IRM, la source est placée arbitrairement en (0,0,0). La quantité étudiée est la fonction de transfert définie par:

$$H \propto \left| \sqrt{\frac{W_{rad}}{Q}} \right|$$

où W_{rad} est la puissance acoustique totale rayonnée aux lèvres.

Le premier cas concerne la voyelle /a/ simulée au moyen de 15 guides élémentaires comme montré sur la figure 5, dans ce cas il a été supposé que la structure est symétrique par rapport à l'axe z.

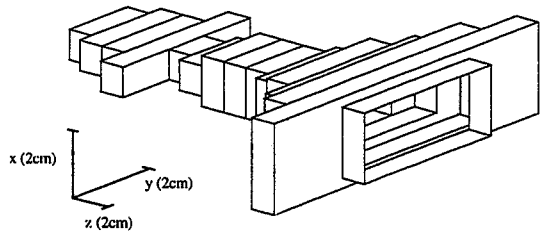


Figure 5: Approximation à 15 sections de la voyelle /a/.

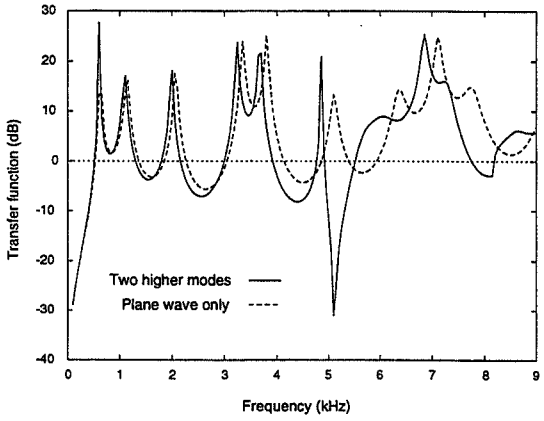


Figure 6 : Fonction de transfert de la voyelle /a/. Comparaison entre une solution unidimensionnelle (onde plane uniquement) et une solution prenant en compte deux modes.

Le second exemple concerne la fricative /ʃ/ approximée par une succession de 38 guides comme décrit sur la figure 7. La fonction de transfert correspondante est présentée sur la figure 8.

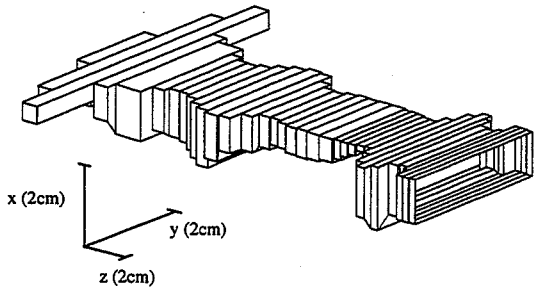


Figure 7: Approximation à 38 tubes de la fricative /ʃ/.

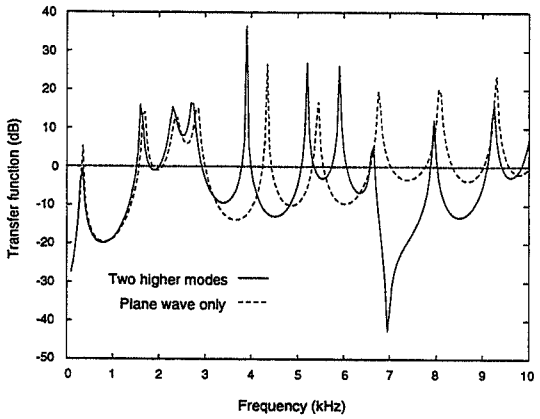


Figure 8 : Fonction de transfert de la fricative /ʃ/.

Les résultats présentés conduisent aux conclusions suivantes :

En dessous de la première fréquence de coupure, la perturbation induite par les modes d'ordre supérieurs est faible : un abaissement des fréquences de résonances de l'ordre de quelques pourcents. A partir de la fréquence de résonance on peut observer un « zéro » puis une forte perturbation des résonances liée à la présence de modes d'ordre supérieurs propagatifs. La position de ces « zéros » étant fonction de la dimension transverse du conduit vocal, elle est donc différente selon le son considéré (5kHz pour le /a/, 7 kHz pour le /ʃ/). Comme on pouvait s'y attendre l'influence des modes d'ordre supérieurs est plus importante dans une configuration asymétrique du conduit vocal. Cependant, on peut constater au vu de la figure 6 que même pour une approximation symétrique l'influence des modes de propagation est clairement visible.

Enfin, il faut noter que les résultats présentés se réfèrent à une source unique et ponctuelle. Une telle approximation, si elle est acceptable au premier ordre pour les sons voisés, ne l'est certainement plus dans le cas des fricatives.

5. CONCLUSION

Les exemples développés montrent que le champ de pression acoustique aux fréquences élevées est fortement déterminé par la présence de modes d'ordre supérieurs. Les effets observés sont à rapprocher des observations

faites sur la parole naturelle : présence de zéros dans la fonction de transfert des voyelles orales ou des fricatives, amplification du contenu spectral du champ rayonné aux fréquence élevées pour les fricatives.

RÉFÉRENCES

[Bru98] Bruneau M.. (1998). Manuel d'Acoustique Fondamentale, Hermes, Paris.

[Elm98] Elmasri S., Pelorson X., Saguet P., Badin P. (1998). "The use of the Transmisson Line Matrix in acoustics and in Speech", International Journal of Numerical Modelling, 11, 133-151.

[Fan60] Fant G. (1970) "Acoustic theory of speech production", 2nd Ed., Mouton and co, Den Haag.

[Ker91] Kergomard, J. (1991) "Calculation of discontinuities in waveguides using mode-matching method : an alternative to the scattering matrix approach", J. Acoustique, 4, 111-138.

Mesures électroglottographiques du quotient d'ouverture glottique en voix parlée et chantée

Nathalie Henrich, Christophe d'Alessandro, Michèle Castellengo, Boris Doval

LAM (UPMC, CNRS, Ministère de la culture), 11 rue de Lourmel, 75015 Paris.
LIMSI-CNRS, BP 133, F91403 Orsay.
E-mail : henrich@lam.jussieu.fr, {cda,doval} @ limsi.fr, castel@ccr.jussieu.fr

ABSTRACT

Measurements of the voice open quotient in speech and singing are reported and discussed. The open quotient is measured using electroglottography, for 2 singers (1 female, 1 male), under various speech and singing conditions (spoken and sung sentences, sustained vowels, crescendo and decrescendo, glissando). The results show : 1/ systematic differences between male and female subjects (the open quotient is lower for males); 2/ strong variation of O_q at a change of laryngeal mechanism, during glissandos; 3/ a singer-dependent correlation between vocal intensity and O_q in singing; 4/ systematic differences between speech and singing (less O_q variations in singing than in speech).

1. INTRODUCTION

La qualité vocale, dans la parole comme dans le chant, repose en grande partie sur les propriétés de la source de voisement. En s'appuyant sur le modèle acoustique source/filtre ([3]), on peut étudier les paramètres de la source de débit glottique pour caractériser les différentes émissions. Plusieurs modèles paramétriques de l'onde de débit glottique ont été proposés ([8], [4], [10], [6]). Ils ne possèdent pas tous les mêmes paramètres, ni le même nombre de paramètres. Néanmoins, dans tous les modèles on retrouve comme paramètres la période de voisement (T_0), le quotient d'ouverture (O_q), et au moins un paramètre qui règle la vitesse de fermeture. Ce quotient O_q est ici défini comme le temps ouvert relatif, c'est-à-dire le rapport de la période ouverte de la glotte sur la période de voisement (et non pas le rapport de la période d'ouverture de la glotte sur la période de voisement, définition que l'on rencontre parfois). Du point de vue spectral [2], il semble bien que l'on puisse distinguer l'effet des paramètres de la source. Le quotient d'ouverture semble lié aux propriétés du signal en basses fréquences, en particulier sur les premiers harmoniques. La vitesse de fermeture influence plutôt la pente spectrale pour les fréquences supérieures à environ 0.5-1 kHz. Du point de vue physiologique, le quotient d'ouverture est lié à la qualité serrée, tendue (O_q petit) , ou détendue (O_q élevé) de la voix. Le quotient d'ouverture est donc un paramètre significatif de la qualité vocale perçue, et il est important de l'étudier pour l'analyse de la qualité vocale et la synthèse de la parole et du chant.

Les questions abordées dans cet article concernent les valeurs et les variations du quotient d'ouverture. Par exemple, quelle est la dynamique du quotient d'ouverture en parole et en chant ? Sont elles comparables ? Peut-on observer des différences entre voix d'homme et voix de femme ? Et selon différents mécanismes laryngés ? Peut-on caractériser la voix chantée en terme de quotient d'ouverture ?

La section 2 décrit le procédé de mesure du quotient d'ouverture, et la base de données analysée. La section 3 présente et discute les résultats obtenus. La section 4 conclue l'étude.

2. MÉTHODE

2.1. Sujets

Deux sujets ont été enregistrés pour le moment, tous les deux professionels du chant et formateurs de la voix. Le sujet B1 est un baryton, le sujet S1 est une soprano.

2.2. Dispositif d'enregistrement

La base de données contient actuellement les enregistrements simultanés du signal acoustique et du signal électroglottographique d'un chanteur et d'une chanteuse. La chaîne de mesure du signal acoustique est constituée d' un microphone de pression 1/2" (Brüel & Kjær 4165) placé à 50 cm du chanteur, un préamplificateur (Brüel & Kjær 2669) et un amplificateur de mesure (Brüel & Kjær NEXUS 2690). Le signal est numérisé directement sur un DAT (PORTA-DAT PDR1000) à une fréquence d'échantillonnage de 44,1 kHz sur 16 bits. Le signal électroglottographique est enregistré simultanément sur la seconde piste du DAT, à l'aide d'un électroglottographe à deux voies (EG2). Les enregistrements sont réalisés dans une cabine insonorisée (mais non anéchoïque). Lors d'une calibration préalable, le chanteur émet un son tenu, dont l'intensité acoustique est mesurée à l'aide d'un sonomètre analogique, placé au niveau du microphone.

2.3. Protocole d'enregistrement

Pendant une séance d'enregistrement, on demande aux sujets de chanter avec le moins de vibrato possible, pour différentes conditions de voix parlée et chantée, dont les suivantes seront exploitées dans cette étude :

voix parlée/voix chantée Une phrase en français, choisie par le sujet, est énoncée, chantée, puis énoncée à nouveau et criée.

phrase musicale les premières mesures de l' *Ave Maria* de Gounod sont interpétées en variant l'intensité et la qualité vocale.

Sons crescendo et voyelles tenues Les trois voyelles [a], [e] et [u] sont émises sur différentes hauteurs et pour différentes intensités. L'attention est portée sur la couleur de la voyelle et la hauteur. Les sons produits ont une durée de 4 à 8 s, et, en le précisant, les chanteurs peuvent éventuellement utiliser l'un ou l'autre des deux mécanismes laryngés.

Glissandos Des glissandos ascendant et descendant, mezzo-forte, continu et de préférence avec le moins de vibrato possible, sont demandés en fin d'enregistrement.

2.4. *Méthode de mesure du quotient d'ouverture*

Le signal électroglottographique, proportionnel à l'impédance électrique du cou, varie en fonction de la surface de contact des cordes vocales. En le dérivant, on obtient pour chaque période deux pics assez nets, de sens opposé. Le pic le plus marqué correspond au moment de fermeture glottique, tandis que l'autre pic peut être relié à l'instant d'ouverture glottique. La détermination de ces deux instants permet alors de connaître la valeur du quotient d'ouverture (cf. figure 1).

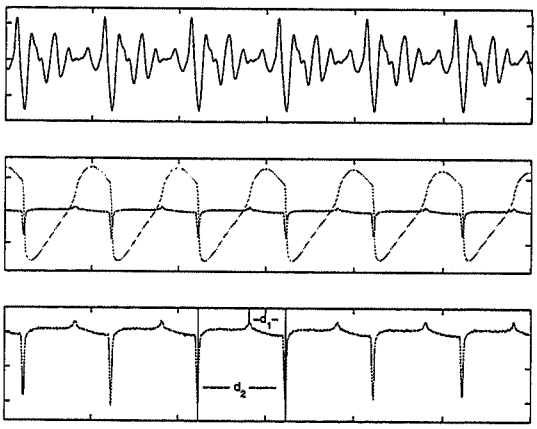


Figure 1: Illustration de la mesure du quotient d'ouverture O_q sur des signaux électroglottographiques (EGG). En haut : le signal acoustique, au milieu le signal EGG et sa dérivée, en bas la dérivée de l'EGG. $O_q = \frac{d_1}{d_2}$

Il faut noter que cette méthode a des limites. On observe parfois une indétermination du pic associé à l'ouverture glottique, ce qui se traduit par des valeurs abérrantes de O_q.

3. RÉSULTATS ET DISCUSSIONS

3.1. *Résultats*

Le tableau 1 montre les valeurs (moyenne et écart type) du quotient d'ouverture mesurées sur les 2 sujets pour différentes conditions. La phrase parlée choisie par le sujet B1 est : "*Qui veut chasser une migraine n'a qu'à boire toujours du bon*"; et celle choisie par le sujet S1 est : "*Adieu notre petite table*". La voyelle présentée ici est le [a], chantée sur un Do 3 (260 Hz) par le sujet B1 et un Do 4 (520 Hz) par le sujet S1.

Table 1: Mesure du quotient d'ouverture pour différentes productions vocales

		B1		S1	
		O_q	σ	O_q	σ
phrase	chantée	0.54	0.04	0.75	0.04
	parlée	0.59	0.09	0.69	0.08
	criée	0.49	0.08	0.69	0.07
Ave Maria	piano	0.63	0.04	0.69	0.03
	médium	0.57	0.05	0.68	0.02
	forte	0.52	0.03	0.70	0.04
Voyelles tenues	piano	0.61	0.02	0.61	0.03
	médium	0.57	0.02	0.67	0.02
	forte	0.50	0.01	0.69	0.01

On remarque sur ces résultats que l'homme et la femme ne partagent pas les mêmes plages de variation de O_q (valeur moyenne) : on observe une variation de 0.49 à 0.63 pour le sujet B1 et 0.61 à 0.75 pour le sujet S1. Ceci est en accord avec les résultats montrant que les quotients d'ouverture sont en moyenne plus élevés chez la femme que chez l'homme ([6], [5], [7]).

3.2. *Effet des mécanismes laryngés*

Cependant, cette différence de valeurs moyennes de O_q s'explique aussi par le fait que la plupart des exemples sont des productions chantées, pour lesquelles les deux sujets n'utilisent pas le même mécanisme vibratoire ([7]). Le baryton utilise principalement le mode 1, pour lequel les cordes vocales sont épaisses et vibrent sur toute leur longueur, tandis que la soprano chante en mode 2, ce qui se traduit physiologiquement par des cordes vocales plus fines et une réduction de la masse vibrante ([9]). Un chanteur n'aura donc pas le même quotient d'ouverture, selon qu'il utilise l'un ou l'autre de ces deux modes. Pour illustrer ce phénomène, la figure 2 représente l'évolution du quotient d'ouverture pour le début de phrase "*qui veut*" chanté en mode 1 (Sol 2, $F_0 = 196$ Hz), chanté en mode 2 (Ré 3, $F_0 = 293$ Hz) et crié en mode 1 ($F_0 = 200$ Hz). On constate que O_q est nettement plus élevé pour le mode 2 que pour le mode 1 et la voix criée.

Cet effet des mécanismes laryngés est également très marqué sur les glissandos produits par les deux chanteurs, au moment du passage d'un mode à l'autre. La figure 3 montre les variations de O_q au cours d'un glissando ascendant et descendant effectué par le sujet B1.

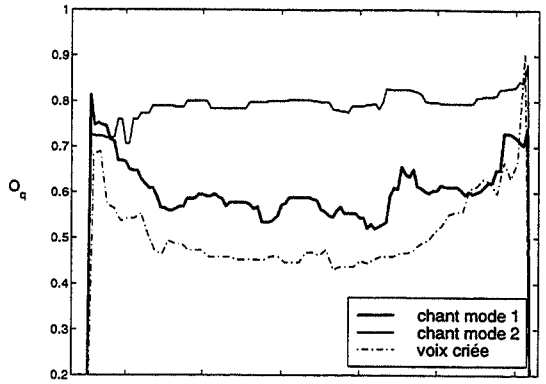


Figure 2: phrase "*qui veut*" chanté en mode 1, en mode 2 et en voix criée par le sujet B1.

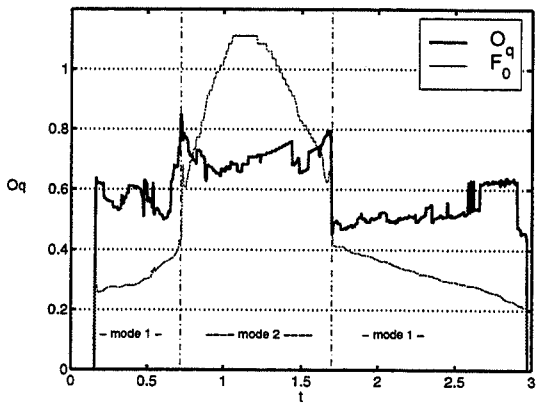


Figure 3: Variations de F_0 (de 200 à 600 Hz) et de O_q lors d'un glissando avec passage de mode 1 à mode 2 et retour au mode 1, effectué par le sujet B1. A noter les décrochements de quotient d'ouverture

Sur cette figure, les variations de O_q et de F_0 mesurées sur le signal EGG ont été superposées. On observe deux décrochements en fréquence au cours du glissando, qui correspondent à deux passages : mode 1 / mode 2 pendant la montée et mode 2 / mode 1 pendant la descente. Ce décrochement en fréquence s'accompagne d'un décrochement de O_q : augmentation subite pour le passage mode 1 / mode 2 et diminution subite pour le passage mode 2 / mode 1.

3.3. Effet de la force de voix

Pour étudier les variations du quotient d'ouverture en fonction de la force de voix, l'intensité du son a été calculée à chaque instant et comparée à O_q. L'intensité est mesurée sur le signal acoustique par calcul d'une énergie à court terme rapportée à la valeur de calibration obtenue par le sonomètre.

La figure 4 représente un exemple de crescendo/decrescendo sur une note fixe. Elle montre clairement le lien entre la force de voix et le quotient d'ouverture.

Une étude systématique des variations du quotient d'ouverture en fonction de la force de voix fait apparaître des différences notables entre les 2 chanteurs. Les figures 5 et 6 montrent la répartition des valeurs de quotient d'ouverture et d'intensité pour la voyelle

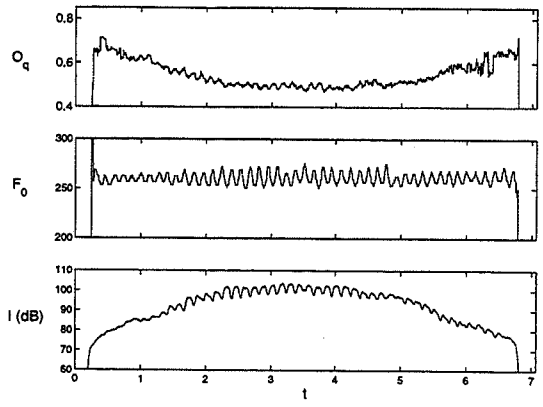


Figure 4: Crescendo/decrescendo du chanteur B1 sur la note Do 3 (260 Hz). En haut : O_q; au milieu : la fréquence fondamentale; en bas : l'intensité mesurée sur le signal acoustique. A noter la covariation de O_q avec l'intensité et la présence de vibrato de quotient d'ouverture.

tenue [a] sur la note Do 3 (baryton) et Do 4 (soprano). La tendance du baryton est de diminuer son quotient d'ouverture à mesure que l'intensité produite augmente. Celle de la soprano est inverse : elle tend à augmenter son quotient en fonction de l'intensité.

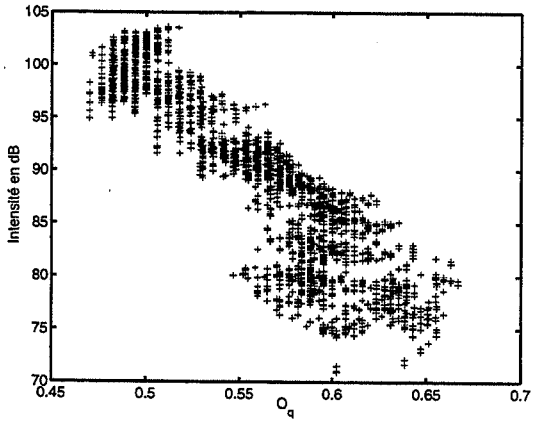


Figure 5: Répartition des valeurs d'intensité en fonction de O_q pour le chanteur B1, sur la note Do 3 (260 Hz)

Si l'une des façons d'augmenter le niveau du son émis est de diminuer le quotient d'ouverture (en "serrant" la voix), ce n'est pas la seule. Les résultats précédent semblent indiquer que le chanteur B1 l'utilise mais pas la chanteuse S1.

3.4. Voix parlée / voix chantée

Nos résultats montrent une différence notable de dynamique du quotient d'ouverture entre les conditions voix parlée / voix chantée, qui se traduit par un écart-type nettement plus important (environ le double) pour la phrase parlée que pour la phrase chantée (cf. table 1). Les grandes et rapides variations observées en voix parlée ne se retouvent pas ou très peu dans le chant.

Cette dynamique du O_q sur la parole se traduit par des variations à l'échelle de temps du phonème. No-

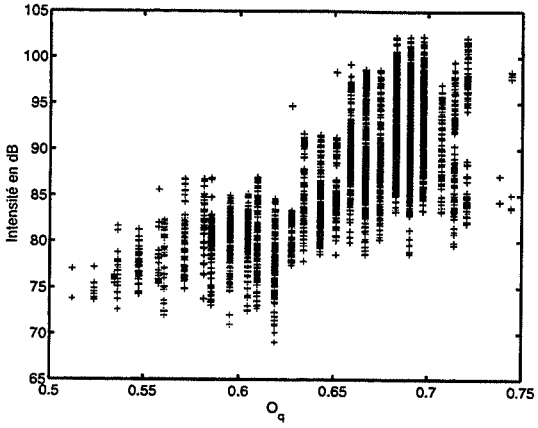


Figure 6: Répartition des valeurs d'intensité en fonction de O_q pour la chanteuse S1, sur la note Do 4 (520 Hz)

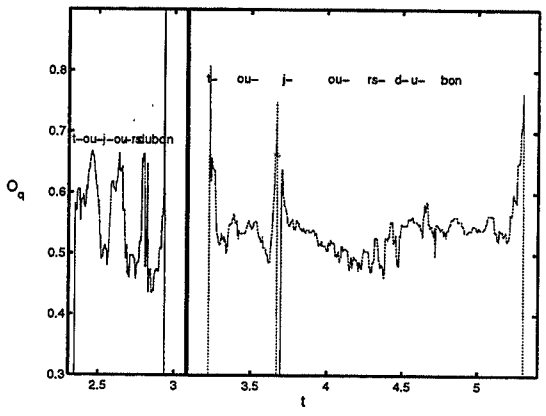


Figure 7: Evolution de O_q pour la fin de phrase "... toujours du bon". A gauche : voix parlée, à droite : voix chantée. Observables sur les occlusives, les variations de O_q en voix parlée ne se retrouvent pas en voix chantée.

tamment, on observe souvent une diminution nette de O_q sur les occlusives (cf figure 7) et aussi une augmentation de O_q sur les débuts et fins de phrase. Il s'agit donc probablement de variations liées à la prosodie et à l'articulation.

Sur la voix chantée, ces variations ne s'observent quasiment pas. C'est un peu comme si les chanteurs avaient appris à gommer les perturbations dues à l'articulation des phonèmes dans le souci d'une plus grande continuité sonore. Il est ainsi remarquable que même pour des variations importantes de F_0 le quotient d'ouverture puisse rester très constant pour la voix chantée.

4. Conclusions

Nous avons présenté les premiers résultats d'une recherche en cours sur la caractérisation des paramètres de la source glottique en voix parlée et chantée. Les mesures de quotient d'ouverture et de fréquence fondamentale sont obtenues par électroglottographie, pour 2 sujets chanteurs. Les résultats principaux de cette étude peuvent se résumer ainsi :

1. des différences systématiques ont été observées entre la voix d'homme et la voix de femme. Les quotients d'ouverture sont plus faibles pour l'homme, en voix parlée comme en voix chantée. Ces différences sont principalement dues aux mécanismes laryngés, même en voix parlée.
2. l'expérience du glissando montre une rupture de O_q au moment du changement de mécanisme laryngé, qui coïncide avec celle de F_0.
3. pour un même chanteur, la force de voix et le quotient d'ouverture sont corrélés, mais le sens de cette corrélation dépend fortement du chanteur. Une étude sur un grand nombre de sujets permettrait de déterminer si cela provient plutôt de différences de genre, de mécanisme laryngé ou de technique vocale.
4. la dynamique du quotient d'ouverture diffère de façon fondamentale entre la voix parlée et la voix chantée. En parole, les variations de quotient d'ouverture sont liées à l'articulation et à la prosodie. En chant, la recherche de continuité sonore semble se traduire par de faibles variations locales de ce quotient.

Ces travaux vont se poursuivre dans plusieurs directions. La base de données va être augmentée de nouveaux sujets. Le paramètre de pente spectrale, qui semble caractériser la force de voix doit aussi faire l'objet d'une étude systématique.

Bibliographie

[1] D.G. Childer and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *J. Acous. Soc. Am.*, 90:2394–2410, 1991.

[2] B. Doval and C. d'Alessandro. Spectral correlates of glottal waveform models : an analytic study. *Proc. ICASSP'97*, pages 1295–1298.

[3] G. Fant. *Acoustic theory of speech production.* Mouton, La Hague, 1960.

[4] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 85(2):1–13, 1985.

[5] H.M. Hanson. Glottal characteristics of female speakers : Acoustic correlates. *J. Acous. Soc. Am.*, 101:466–481, 1997.

[6] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acous. Soc. Am.*, 87(2):820–857, 1990.

[7] F. Lecluse and M. Brocaar. Quantitative measurements in the electroglottogram. *17th International Congress of Logopedics and Phoniatrics*, 1977.

[8] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J. Acous. Soc. Am.*, 49:583–590, 1971.

[9] B. Roubeau and M. Castellengo. Revision of the notion of voice register. *XIXth International CoMeT Congress, Utrecht*, 1993.

[10] R. Veldhuis. A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *J. Acous. Soc. Am.*, 103:566–571, 1998.

Détermination de la position du voile du palais à partir du signal de parole pour les nasales du français.

Solange Rossato, Pierre Badin, Gang Feng

Institut de la Communication Parlée
UMR 5009, CNRS,Univ. Stendhal, INPG – Grenoble, France
Tél.: ++33 (0)476 82 41 20 - Fax: ++33 (0)476 82 38 45
Mél: rossato@icp.inpg.fr

ABSTRACT

This paper deals with the recovery of the velar position from speech signals for the French nasals for one subject. Articulatory movements were recorded using electromagnetic articulography (EMA) synchronized with the acoustic signals. The maximum of likelihood method was used to estimate the velar position from speech signals. The results show that the velum position is well recovered with a root mean square error of 0.12 cm, when the learning is performed over the complete corpus. Higher precision is achieved when nasal vowels and nasal consonants are considered separately.

1. INTRODUCTION

L'inversion de la parole consiste à retrouver les mouvements articulatoires à partir du signal de parole. Ce thème de recherche a donné lieu à de nombreux travaux ces dernières années. Les enjeux sont grands autant pour le codage à bas débit et la synthèse articulatoire que pour l'aide à la rééducation. Retrouver les gestes articulatoires à partir du son est un problème mal défini dans le sens où plusieurs configurations articulatoires peuvent donner des signaux acoustiques très proches ([Abr94]). De nombreuses approches ont besoin de connaissances préalables issues de données, par exemple pour construire un dictionnaire de liens entre paramètres acoustiques et paramètres articulatoires, pour entraîner un réseau de neurones.

Les relations entre espace acoustique et espace articulatoire sont complexes, et l'effet d'un articulateur sur le signal acoustique n'est pas toujours bien connu. Dans le cas particulier de la nasalité, un deuxième conduit, le conduit nasal, entre en jeu et ajoute une complexité supplémentaire. Ainsi, si les voyelles sont bien caractérisées par leurs formants, la nasalisation d'une voyelle introduit des pôles et des zéros qui ont des effets différents suivant la voyelle et le degré de couplage entre conduit oral et conduit nasal ([Che95], [Fen96]). Le principal effet de ces paires de pôles / zéros est un aplatissement du spectre en basse fréquence, un élargissement des bandes passantes, et une atténuation des formants qui deviennent ainsi plus difficiles à détecter.

La grande variabilité du signal de parole et la difficulté d'observations des mouvements du voile du palais nous a conduit à réaliser une première étude ([Ros98]) utilisant une modélisation des voyelles nasales. L'avantage de la modélisation réside dans le contrôle tous les paramètres ainsi que dans une grande facilité pour générer des données. L'utilisation du maximum de vraisemblance a montré des résultats encourageants avec un taux d'estimation correcte d'environ 90%. Le travail présenté dans cet article vise maintenant à tester la méthode du maximum de vraisemblance non plus sur des données issues d'un modèle mais sur des données enregistrées sur un locuteur homme avec un articulographe électromagnétique (EMA).

2. LES DONNEES

Deux corpus ont été enregistrés pour un seul locuteur homme, l'un pour les voyelles nasales, l'autre pour les consonnes nasales. L'objectif est d'obtenir des articulations ne se différenciant que pour le voile du palais. Le locuteur prononce plusieurs répétitions des quatre voyelles nasales du français /ɑ̃/, /ɔ̃/, /ɛ̃/ et /~œ/ /, suivies de leur contrepartie orale avec pour consigne de ne pas bouger les articulateurs autres que le voile du palais. Dans le cas des consonnes, des séquences /pVNVpVCV/ ont été enregistrées, où N représente la consonne nasale et C la consonne occlusive correspondante, en contexte /a i u/. Les paires de consonnes nasale/orale ont été choisies sur le critère d'une articulation proche où seule la position du velum diffère : paires n/d et m/b. Le corpus des voyelles est constitué de 36 réalisations et celui des consonnes en contient 102.

2.1 Description du dispositif expérimental

Deux capteurs EMA ont été fixés au niveau des incisives supérieures (référence) et des incisives inférieures (mâchoire). Deux autres capteurs ont été positionnés sur la langue (pointe, partie laminaire). Un cinquième capteur a été collé sur le velum (voir figure 1).

Les coordonnées des capteurs sont échantillonnées à 1 kHz. Les mouvements des articulateurs sont relativement lents ce qui permet d'appliquer un filtre passe-bas sur toutes les données afin d'éliminer une partie du bruit de mesure. Un micro permet d'enregistrer le signal de parole sur DAT avec un excellent rapport signal à bruit. Ce signal est ensuite échantillonné à 16 kHz et les trames sont synchronisées à 100 Hz avec les positions des capteurs.

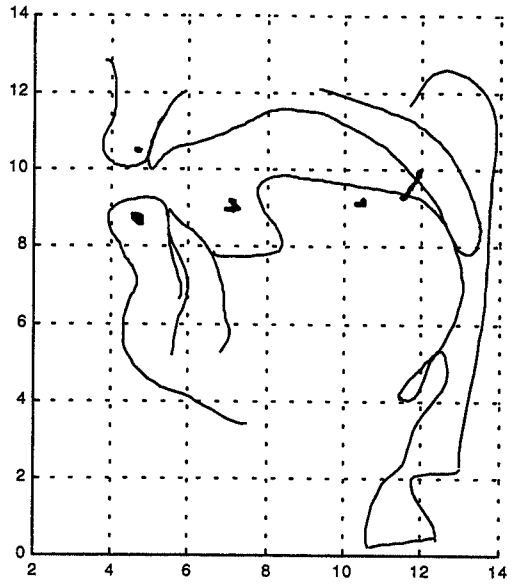


Figure 1. Trajectoires des cinq capteurs EMA pour la transition /ãa/ superposées avec le contour obtenu par IRM pour le même sujet et pour l'articulation /ã/.

Les trajectoires des capteurs permettent de vérifier les mouvements des articulateurs du locuteur. Concernant le corpus des voyelles nasales et de leur contrepartie orale, les capteurs indiquent un très léger mouvement de la langue lors de la transition (de l'ordre de 1 mm) et aucun mouvement de la mâchoire. L'observation des capteurs pour le corpus des consonnes montre que la durée moyenne de l'occlusion dentale est du même ordre de grandeur pour le /n/ que pour le /d/ avec des évolutions semblables pour les articulateurs autres que le velum. Le critère de départ qui consiste à obtenir des articulations ne se différenciant que pour la position du velum semble donc respecté.

2.2 Traitement des signaux

Paramètre indicateur de la position du voile du palais Le but de ce prétraitement est de déterminer un paramètre unique qui soit représentatif des mouvements du velum au lieu des deux coordonnées X et Y du capteur EMA du velum. Le mouvement de ce capteur est relativement rectiligne avec une pente qui varie peu d'un item à l'autre. Le choix du paramètre « hauteur » du velum se porte donc vers la projection du point (X,Y) sur une droite. La droite choisie est la moyenne des droites de régression obtenues pour chaque item.

Cette " hauteur " du velum (VH) a un mouvement d'amplitude 0,8 à 1 cm environ pour les voyelles nasales / ɛ̃ / et /~œ/ ; plutôt entre 0,6 et 0,8 cm pour les voyelles plus ouvertes /ã/ et /õ/. Pour les séquences /pVNVpVCV/, les mouvements sont en général d'amplitude plus faible. La figure 2 présente deux exemples de courbes obtenues pour le paramètre VH.

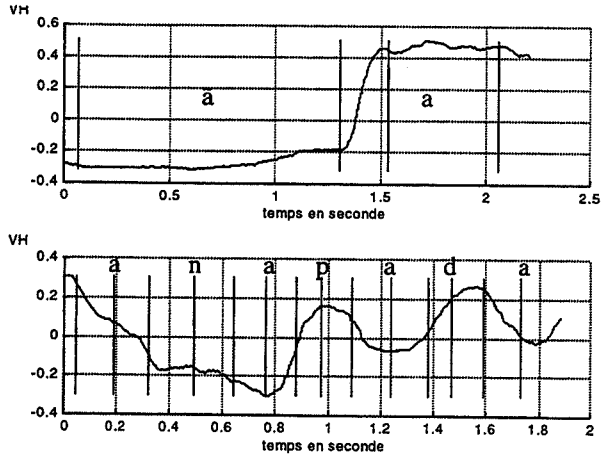


Figure2: Paramètre VH " hauteur " du velum en cm. Pour l'item /ãa/, l'amplitude du mouvement est de l'ordre de 0,8 cm. Pour l'item /panapada/, on remarque trois niveaux : relevé pour /p/ et /d/, abaissé pour /n/, avec une position intermédiaire pour les /a/. La voyelle /a/ qui suit /n/ garde la position très basse du velum atteinte pendant la consonne /n/. Ce phénomène est également observé pour les voyelles /i/ et /u/ suivant la consonne /n/.

Signaux acoustiques Dans un premier temps, le signal acoustique est étiquetté pour faciliter l'extraction d'une portion du signal. Le signal acoustique est découpé en trames de longueur 32 ms avec un recouvrement de 22 ms. Pour chaque trame, le signal est d'abord filtré passe-haut puis une analyse par banc de filtre fournit 16 valeurs à partir desquelles on obtient les 16 coefficients melcepstres (MFCC).

3. MAXIMUM DE VRAISEMBLANCE

3.1 La Méthode

Le signal acoustique, représenté par un vecteur S de coefficients, dépend de façon complexe du paramètre VH. Il s'agit de décider, à partir d'un vecteur S, quelle est la valeur la plus probable de VH, notée VH_{est}. La méthode du maximum de vraisemblance peut s'appliquer à ce problème : la valeur de VH_{est} est celle qui maximise la vraisemblance du vecteur S pour le paramètre VH, p(S/VH). Puisque VH est un paramètre continu, la probabilité que le paramètre estimé soit une valeur précise VH_0 est très faible : une certaine incertitude existe autour de cette valeur. L'ensemble des valeurs que peut prendre VH est donc découpé en N intervalles. Si les intervalles sont trop petits, les probabilités sont plus faibles et les résultats moins fiables. Si les intervalles sont trop grands, la précision est plus faible. Il faut donc trouver un compromis entre fiabilité et précision dans l'estimation.

Connaissant le vecteur S, on calcule, pour chaque intervalle V_i de centre v_i, la probabilité conditionnelle $p(S/v_i)$ aussi appelée vraisemblance du vecteur S sachant

Tableau 1 : Le tableau 1 présente les erreurs RMS (en cm) obtenues pour les voyelles nasales ($\tilde{V}$), les transitions (tr) et leur contrepartie orale (V) ainsi que pour les consonnes nasales (N) et orales (C) et les voyelles adjacentes : VNV et VCV, pour différentes bases d'apprentissage.

Apprentissage	$\tilde{V}$	tr	V	V	N	V	V	C	V
Les 2 corpus	0.08	0.22	0.10	0.10	0.11	0.11	0.08	0.08	0.10
Corpus voyelles	0.09	0.14	0.09	-	-	-	-	-	-
Parties vocaliques	0.07	0.20	0.12	0.07	0.23	0.12	0.07	0.08	0.08
Consonnes	-	-	-	-	0.08	-	-	0.07	-

que le paramètre VH a pour valeur v_i. Pour un vecteur S donné, on obtient une probabilité pour chaque intervalle de centre v_i. L'estimation du paramètre VH est la valeur v_i qui a la plus grande probabilité. On peut également calculer l'espérance en associant à chaque centre v_i, la probabilité $p(S/v_i)$ normalisée. Le paramètre estimé VH_{est} est alors un paramètre continu comme l'est le paramètre VH, « hauteur » du velum.

3.2 Détermination des probabilités conditionnelles

Pour appliquer la méthode du maximum de vraisemblance, les probabilités conditionnelles $p(S/v_i)$ doivent être connues au préalable. Ces probabilités sont considérées comme des lois normales estimées à partir de statistiques effectuées sur les données articulatoires et acoustiques extraites des corpus enregistrés : ensemble de vecteurs de coefficients MFCC et valeur du paramètre VH correspondant. La plage de variation de VH est divisée en intervalles de largeur 0,1 cm. Cette taille permet d'avoir assez de données pour chaque intervalle pour déterminer la probabilité conditionnelle $p(S/v_i)$: entre 300 et 5000 vecteurs S sur lesquels sont calculées moyenne et matrice de covariance. Les lois normales sont donc entièrement définies. Pour chaque corpus, entre 9 et 12 itérations de chaque item sont prononcées par le locuteur. Les 5 premières sont utilisées pour l'apprentissage, c'est-à-dire pour déterminer les probabilités conditionnelles, les autres sont réservées au test.

4. ANALYSE DES RESULTATS

La détermination de la position du voile du palais est effectuée tout d'abord de façon globale en prenant en compte sans distinction le corpus de voyelles et le corpus de consonnes. Ensuite, une étude restreinte au corpus des voyelles est présentée avant d'étendre l'apprentissage à toutes les parties vocaliques des corpus. Enfin, nous nous intéressons à la détermination de la position du velum lors de la production de consonnes.

4.1 Analyse globale

L'apprentissage est effectué « en aveugle » sur 100 000 trames provenant des deux corpus. La variation maximale de la « hauteur » du velum est de 1,17 cm. 11 intervalles égaux de largeur 0.1 cm sont utilisés avec plus de 1000 trames par intervalle. L'erreur quadratique moyenne (RMS) entre la « hauteur » déterminée par la méthode du maximum de vraisemblance VH_{est} et la « hauteur » mesurée VH est de 0.12 cm pour l'ensemble du corpus de test. Cette valeur, si elle donne une idée globale de l'estimation, ne permet pas de d'établir un lien entre qualité de la détermination de la position du velum et phonème. Les erreurs RMS entre la détermination de VH_{est} et VH sont présentées de façon détaillée dans le tableau 1. Dans l'ensemble, le paramètre VH est déterminé avec une précision correcte. Cependant, le paramètre VH_{est} est déterminé avec une erreur RMS importante (0,22 cm) pour les transitions des voyelles nasales vers leur contrepartie orale.

4.2 Voyelles nasales et parties vocaliques

Un premier apprentissage a été réalisé sur un corpus limité contenant les quatre voyelles nasales, leurs contreparties orales et les transitions, soit environ 38 000 trames de signal acoustique. L'erreur RMS est de 0.08 cm pour les voyelles nasales et leur contrepartie orale. Pour les transitions, l'erreur RMS de détermination de la position du velum est de 0,14 cm. Cependant, le coefficient de corrélation moyen entre VH_{est} et VH pour l'ensemble des transitions est de 0,9. L'évolution est donc relativement bien respectée.

Les voyelles /a/ ,/i/ et /u/ en contexte nasal et non nasal (30 000 trames) ont ensuite été intégrées dans la base d'apprentissage. Les erreurs d'estimation sont présentées dans le tableau 1. On observe une erreur importante (0,20 cm) pour les transitions et une corrélation plus faible (0,84). Dans l'ensemble, la détermination du paramètre VH_{est} est moins précise lorsque les parties vocaliques du corpus des consonnes sont introduites dans la base d'apprentissage. En contrepartie, il est possible de déterminer le paramètre VH_{est} pour les parties vocaliques du corpus des consonnes avec une faible erreur RMS (de l'ordre de 0,07 cm). Seule la voyelle suivant la consonne nasale se démarque avec une erreur de 0,12 cm. D'après les données articulatoires, le velum reste en position basse durant la voyelle qui suit la consonne nasale (voir figures 1 et 3). L'apprentissage pour ces positions de velum se fait principalement grâce aux transitions des voyelles nasales vers leur contrepartie orale. On retrouve donc des erreurs supérieures, plus proches de celle obtenues pour les transitions.

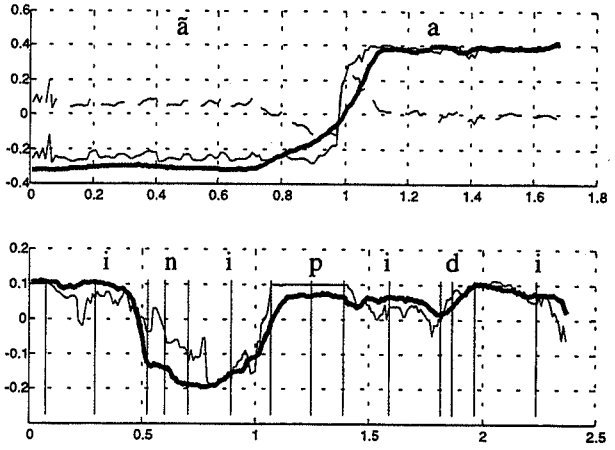


Figure 3. Exemples de VH_{est} et de VH pour l'item /ãa/ (en haut) et l'item /pinipidi/ (en bas). Le trait fin représente VH_{est} tandis que le trait épais correspond à VH. La différence a été tracée (pointillé) pour l'item /ãa/ : les erreurs sont importantes pour la transition.

4.3 Consonnes nasales

Un apprentissage spécifique aux consonnes nasales et orales a été réalisé sur 14 000 trames. La « hauteur » du velum a une plage de variation plus faible (0,8 cm) sur la réalisation des consonnes que sur celle des voyelles (1,1 cm). L'erreur RMS commise lors de la détermination de la « hauteur » du velum est de 0,08 cm pour les consonnes nasales /n/ et /m/ et de 0,07 cm pour les consonnes orales /d/ et /b/.

5. Discussion et perspectives

En conclusion, la méthode du maximum de vraisemblance permet de déterminer le paramètre VH_{est} à partir du signal acoustique avec une erreur de 0,12 cm avec un apprentissage global. Un apprentissage plus ciblé permet d'obtenir des erreurs légèrement inférieures. Les modifications du signal acoustique introduites par l'abaissement du voile du palais semblent permettre de « séparer » les vecteurs acoustiques indépendamment de la qualité de la voyelle.

Ces résultats peuvent être comparé à ceux de [Ric99] qui utilise un réseau de neurones pour estimer la position du velum. L'apprentissage se fait sur un corpus de 400 phrases, et Richmond trouve des erreurs comprises entre 0,14 et 0,20 en unité arbitraire (proches de nos centimètres) pour les voyelles et les consonnes nasales étudiées. Les erreurs obtenues dans notre étude sont légèrement inférieures pour un apprentissage effectué sur des corpus ciblés sur les voyelles et les consonnes nasales. On pourra envisager dans la suite une pré-détermination de la nature du segment (voyelle / consonne) de manière à pouvoir ensuite utiliser la méthode du maximum de vraisemblance avec la base d'apprentissage appropriée à chaque segment. Cela devrait permettre d'améliorer encore la précision des résultats.

Un point critique de la méthode réside dans le découpage en intervalles. Ce découpage induit une discrétisation arbitraire du paramètre VH. Le nombre d'intervalles est donc un choix important. L'influence de ce découpage sur la détermination de la « hauteur » du velum peut être étudié avec le modèle articulatoire et avec les données mesurées.

La détermination de la position du voile du palais à partir du signal de parole a été appliquée pour un seul locuteur homme il s'agit maintenant d'étendre cette étude à plusieurs locuteurs.

6. Remerciements

Ce travail n'aurait pu se faire sans Pascal Perrier et Christophe Savariaux que nous remercions. Leur aide et leur expérience ont été d'un grand secours pour les mesures articulatoires faites avec l'articulographe

Bibliographie

[Abr94] Abry C., Badin P. et Scully C. (1994) "Sound-to-gesture inversion in speech : The Speech Maps approach", Advanced speech applications pp. 182-196.

[Che95] Chen M.Y. (1995) "Acoustic parameters of nasalized vowels in hearing-imparired and normal-hearing speakers", J.Acoust. Soc. Am. 98 (5), pp. 2443- 2453.

[Fen96] Feng G. et Castelli E. (1996) "Some acoustic features of nasal and nasalised vowels : A target for vowel nasalisation", J. acoust. Soc. Am. 99 (6), pp. 3694-3706.

[Hog96] Hogden J., Löfqvist A., Gracco V., Zlokarnik I. Rubin P. et Saltzman (1996) "Accurate recovery of articulator positions from acoustics : New conclusions based on human data", J. Acoust. Soc. Am. 100 (3), pp. 1819-1834.

[Wre99] Wrench A.A. (1999) "An investigation of sagittal velar movement and its correlation with lip, tongue and jaw movement", Proc. Int. Conf. on Phon. Sc. San Franscisco, pp.435-438.

[Ric99] Richmond K. (1999) "Estimating velum heigth from acoustics during continuous speech", Eurospeech Budapest, pp.149-152.

[Ros98] Rossato S., Feng G. et Laboissière R. (1998) "Recovering gesture from speech signals: a preliminary study for nasal vowels", Proc. ICSLP, Sydney.

Etude aérodynamique de la nasalité en français

Véronique DELVAUX

Aspirant F.N.R.S au Laboratoire de Phonologie
Université Libre de Bruxelles - Belgique
Tél.: +32 2 650 20 18 - Fax: +32 2 650 20 07
Mail: vedelvau@ulb.ac.be - http://www.ulb.ac.be/philo/phonolab

ABSTRACT

This paper presents data about the aerodynamic parameters of French nasality. Data show that the amount of nasalization of a phoneme depends on its own characteristics as well as on its environment. Oral vowels with a phonemic nasal counterpart are less nasalized than the other ones. Regarding oral consonants, the amount of nasalization is greater for fricatives than stops and for voiced than voiceless consonants. Due to tongue configuration, French back nasal vowels [ɔ̃,ɑ̃] have more nasal airflow than [ɛ̃,œ̃]. Context is very powerful in determining the nasalization rate of a segment; carryover nasalization is much stronger than anticipatory nasalization for our 8 Belgian French speakers.

1. INTRODUCTION

Dans cet article, nous présentons les résultats d'une étude des paramètres aérodynamiques de la nasalité en français. Notre point de vue est prioritairement descriptif, même si ce travail s'inscrit dans une recherche plus vaste consacrée à l'implémentation phonétique des traits, et du trait de nasalité en particulier, dans une perspective dynamique. L'objectif de cet article est de présenter, de commenter, et de proposer une explication succinte à un ensemble de résultats concernant le degré de nasalité des segments phonologiquement nasals du français ainsi que des consonnes et des voyelles orales en contexte nasal.

2. MÉTHODE

2.1. Corpus

Cette recherche a été réalisée au Laboratoire de Phonologie de l'Université Libre de Bruxelles auprès de huit sujets belges francophones, quatre hommes et quatre femmes âgés de 22 à 45 ans. Leur tâche était de lire des listes de mots. Le corpus était constitué de 208 items (seuls ou dans de courtes phrases) présentant des séquences nasale-orale de nature diverse (table 1).

Table 1: Présentation schématique du corpus

Séquences	Exemples	Domaine
V̰	un, an, on...	V̰ = ɛ̃,œ̃,ɔ̃,ɑ̃
CV̰C	tente, ponte, pince,...	V̰ = ɛ̃,ɔ̃,ɑ̃; C_1-C_2 = t-t, p-t, p-s.
CV̰.CV	tenter, ton thé,...	V̰ = ɛ̃,œ̃,ɔ̃, ɑ̃ ; C_1=C_2=t ; V=e
CV̰ vs CV	paon-pas, fin-fait,...	C=p,b,f,v,t,d,s,z,ʃ,ʒ,k,g,l,r ; V̰=ɛ̃,ɔ̃,ɑ̃
NV vs NV̰	ma-ment, mot-mon, mou, noeud, nid,...	C=m,n,ɲ ; V̰=ɛ̃,ɔ̃,ɑ̃ ; V=e,i,u,o,y,ø,a,œ,ɔ,ɛ
(C)VN vs V̰N	la nuit-l'ennui,...	V= a,œ,ɔ,ɛ ; N=m,n ; V̰=ɛ̃,œ̃,ɔ̃, ɑ̃
NVN	nonne, manne,...	N_1 et N_2=m,n ; V= i,a,œ,ɔ,ɛ

2.2. Paramètres d'analyse

A l'aide de la station de travail Physiologia [Tes90], nous avons enregistré le signal de parole ainsi que les débits d'air buccal et nasal, puis nous avons déterminé les frontières des segments à étudier avec le logiciel iShell. Pour chacun des segments, cinq paramètres ont été examinés: les débits d'air buccal et nasal moyens (DAB et DAN, en ml/sec), la durée (T, en ms), le volume d'air total sorti par le nez au cours du segment (VAN, en ml) ainsi que la proportion (en %) du débit d'air total imputable au DAN, soit DAN/(DAN+DAB)= PNA (pour "proportional nasal airflow").

A propos du choix de ces paramètres, il faut noter que le débit d'air moyen relevé à la sortie des fosses nasales n'est qu'une mesure indirecte de l'activité vélique. Une variation (1) du débit d'air total ou (2) de la constriction dans le conduit buccal peut déterminer l'évolution au cours du temps d'un tracé de DAN alors que l'ouverture vélique reste constante, ou encore rendre compte de différences moyennes de DAN entre des sons pour lesquels l'abaissement du voile est équivalent [Huf93].

La mesure que nous avons appelée PNA permet de neutraliser l'influence de (1) : si seul le débit d'air total change, c'est-à-dire que l'ouverture vélique et la configuration orale restent constantes, le PNA ne varie pas. Par contre, pour un même débit d'air total, le PNA variera en raison d'un changement de configuration du conduit vocal (2), ainsi que le DAN. Lorsqu'on traite des voyelles nasales entre elles par exemple, il est important de rendre compte de telles variations car il n'est pas exclu qu'elles jouent un rôle au point de vue perceptuel. Il est beaucoup moins utile de comparer des consonnes et des voyelles du point de vue de leur PNA uniquement, puisqu'elles diffèrent complètement au niveau de leur articulation orale. La mesure de PNA présente aussi un intérêt certain, en ceci qu'elle permet de neutraliser les différences quantitatives liées à la variable du sexe du locuteur (voir ci-dessous).

2.3. Stratégie d'analyse

Cette étude se fonde sur des données chiffrées et rompt avec une attitude largement répandue, qui consiste à envisager uniquement les tracés aérodynamiques d'un point de vue qualitatif, comme simples indicateurs de l'activité vélique. L'objectif de cet article est d'analyser la variation de *l'amplitude* des paramètres retenus (DAB et DAN, T, VAN, PNA) en relation avec les conditions de production des sons étudiés : type de segment (nasal ou oral, consonantique ou vocalique), nature du contexte, facteurs externes tels que le sexe du sujet, etc.

Une telle analyse requiert que l'on accorde un soin particulier à deux étapes de l'acquisition des données : la calibration de l'appareil d'une part, et la détermination des frontières des segments à étudier, d'autre part. Ces deux facteurs agissent respectivement sur les axes vertical (quantitatif) et horizontal (temporel) servant traditionnellement de référence aux tracés aérodynamiques, et influencent donc directement les valeurs retenues pour une analyse chiffrée.
On ne perdra pas de vue que les mesures sont des moyennes effectuées sur la totalité d'un segment et qu'elles ne rendent donc pas compte de l'évolution du paramètre aérodynamique au cours du segment. Des événements survenant aux extrémités des segments entrent en compte dans le calcul de cette moyenne, et font en sorte, par exemple, que le DAB reporté dans nos résultats n'est pas nul (cas des consonnes occlusives, table 2-E) alors que la bouche est restée fermée pendant la majeure partie des segments concernés.

2.4. Présentation des résultats

La table 2 présente un résumé des résultats obtenus pour les cinq variables dépendantes étudiées. Huit variables indépendantes ont été sélectionnées; les deux premières concernent l'ensemble des cas analysés, les suivantes ne portent que sur des sous-ensembles, respectivement les voyelles orales (C-D), les consonnes orales (E-F) et les voyelles nasales (G-H). Une analyse de variance (ANOVA) a été effectuée afin de déterminer la significativité des différences moyennes observées entre les groupes relevant d'une même variable. La méthode retenue de calcul de la somme des carrés tient compte des particularités du corpus (taille inégale des cellules, cellules vides). Nous commenterons également au cours de l'analyse qui va suivre, mais sans les présenter en détail faute de place, les résultats les plus saillants des tests complémentaires effectués (tests post hoc avec l'indice de Bonferroni, ANOVA univariée mesurant l'interaction entre certaines variables indépendantes, etc).

3. ANALYSE DES RÉSULTATS

3.1. La variable de sexe (table 2-A)

Les résultats indiquent que le DAN est significativement moins élevé pour les femmes que pour les hommes, tous types de segment confondus (de même que pour chaque paire de type de segment, ainsi que l'ont montré les tests post hoc). Cependant, les femmes ont, par rapport aux hommes, un déficit de DAB proportionnellement bien plus important encore. On peut faire l'hypothèse que cette différence de débit d'air total est due aux différences physiologiques (de capacité pulmonaire) entre hommes et femmes. Par ailleurs, les femmes ont un PNA supérieur aux hommes, en moyenne. Cette différence n'est toutefois pas significative et l'on peut considérer que, pour l'analyse qui va suivre, la variable PNA permet de neutraliser la variation due au sexe du locuteur.

3.2 Le type de segment (table 2-B)

Comme on était en droit de l'attendre, le DAN des segments nasals est nettement supérieur à celui des voyelles et consonnes orales. La distribution de fréquence des valeurs de DAN pour les orales (non présentée ici) montre un pic à une valeur inférieure à la moyenne et un histogramme incliné vers la droite. Pour les voyelles orales, ceci est dû (entre autres) au fait qu'elles sont souvent situées en contexte nasal. Les valeurs de DAN sont donc surestimées en raison de la nature du corpus.
Il n'est pas nécessaire de supposer une ouverture vélique plus importante pour les consonnes que pour les voyelles nasales si l'on veut rendre compte des différences de DAN entre ces deux types de segment. Elles peuvent s'expliquer par la différence de configuration du conduit vocal. Les consonnes nasales sont des occlusives, avec un DAB nul, et donc, à ouverture vélique et pression sous-glottique équivalentes, elles ont un DAN plus élevé. La différence de PNA (plus grande encore) entre N et Ṽ relève des effets conjoints des différences de DAB et de DAN. La valeur de DAB que nous obtenons pour les consonnes nasales (20,2 ml/sec) est due pour [m,n] à des phénomènes de coarticulation lorsque le débit de parole était élevé, et à l'articulation spécifique de [ɲ], à la fin de laquelle nos tracés montrent un relâchement de l'occlusion en une approximante palatale.
Par ailleurs, l'ANOVA a révélé que la différence de DAN (et de PNA) entre voyelles orales et voyelles nasales interagit avec la variable de contexte, alors que ce n'est pas le cas pour les consonnes (C vs N). La figure 1 illustre cette interaction. Deux faits sont particulièrement notables: la différence de PNA entre voyelles orales et nasales est plus importante en contexte oral que nasal; une voyelle précédée d'une consonne nasale présente un plus fort taux de nasalité qu'une voyelle suivie de N. La comparaison des troisième et sixième boxplots de la figure 1 montre qu'une voyelle orale et une voyelle nasale peuvent avoir, en moyenne, un PNA équivalent en raison de l'influence que leur environnement exerce sur elles.

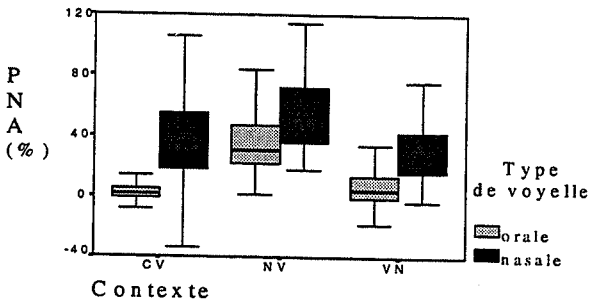


Figure 1 : Interaction des variables de contexte et de type de voyelle pour le PNA

3.3. Les voyelles orales (table 2-C et D)

Comme cela apparait dans la table 2, deux facteurs ont été trouvés qui influencent le DAN des voyelles orales : le contexte et le type de voyelle orale.
Le contexte. Des tests post hoc ont montré que chacun des contextes donnait un DAN significativement différent des autres. On n'a donc pas seulement une opposition contexte oral-contexte nasal, mais également un plus grand taux de nasalité pour les voyelles orales entourées de deux N; la différence la plus remarquable est celle qui oppose encore une fois les voyelles (ici les orales) précédées de N à celles qui sont suivies de N. Nous

reviendrons dans la discussion sur cette différence entre nasalisation progressive et régressive en français.

Le type de voyelle orale. Les données que nous avons recueillies indiquent qu'un processus plus actif pourrait également être à l'oeuvre dans la nasalisation des voyelles orales en contexte consonantique nasal. En effet, une analyse de la variabilité des DAN et PNA selon la voyelle orale considérée a révélé que l'on peut scinder ces voyelles en deux groupes : [i,u,y] et les autres. Les premières (les plus fermées) ont nettement plus de DAN et de PNA que les secondes, qui forment un groupe homogène au sein duquel on ne peut pas lier le taux de nasalité au degré d'aperture de la voyelle. Ces résultats sont en nette contradiction avec diverses études antérieures ([Hou56] et [Mae82]), qui tendent à montrer que les voyelles les plus ouvertes sont les plus nasalisées, et que les fermées sont celles qui ont le moins besoin d'un taux de nasalité élevé pour être perçues comme telles. Les résultats reportés dans la table 2 indiquent que le facteur déterminant, pour les voyelles orales du français, est celui de l'existence ou non d'une contrepartie (phonologique) nasale à la voyelle considérée. Il n'y a plus aujourd'hui en français de voyelles nasales qui correspondent aux voyelles les plus fermées [i,u,y], et ceci pourrait autoriser une nasalité contextuelle plus importante que pour [a,ɛ,ɔ] ou même que les semi-fermées [o,e], qui ont une configuration du conduit vocal fort comparable encore à celle des nasales [ɔ̃,ɛ̃]. Le comportement de [œ] – seule voyelle orale à occuper, dans certaines conditions, une position intermédiaire entre les deux groupes précités, alors que [œ̃] n'est plus un phonème pour quatre de nos sujets – est une indication supplémentaire en faveur d'un *contrôle actif* par le locuteur de l'activité vélique en fonction de la nécessité de préserver (ou non) un contraste maximum entre les deux membres d'une paire phonologique [Kin94].

3.4. Les consonnes orales (table 2-E et F)

Les consonnes orales présentent un certain taux de nasalité. Il ne varie pas selon que la voyelle suivante est nasale ou non. Deux autres variables l'influencent de façon significative. Tout d'abord, les fricatives ont plus de DAN que les occlusives. Ceci parait contraire aux contraintes aérodynamiques qui président à la production des fricatives, pour lesquelles les conditions d'émergence d'une turbulence dans la cavité orale doivent être réunies. Remarquons néanmoins que le débit d'air total est extrêmement élevé pour ces consonnes, précisément afin de créer la turbulence, ce qui induit dans nos résultats un DAB très élevé et relativise quelque peu l'importance du DAN observé. Toutefois, il faudrait pousser plus avant l'investigation dans ce domaine, de même que dans le cas suivant, pour lequel nous ne disposons pas d'une explication satisfaisante: les sonores ont significativement plus de DAN que les sourdes, alors que, produites avec la glotte fermée, elles ont environ deux fois moins de DAB.

3.5. Les voyelles nasales (table 2-G et H)

C'est la combinaison de deux facteurs qui rend compte de la variance du taux de nasalité parmi les voyelles nasales : la nature de la voyelle, et le contexte dans lequelle elle est produite.

Table 2 : Moyennes (arrondies à la 1ère décimale): DAN et DAB (ml/sec), PNA (%), T(ms),VAN(ml). Significativité (ANOVA) pour 8 variables indépendantes.

Variable indépendante	DAN	DAB	PNA	T	VAN
A. Sexe du locuteur					
féminin	32,7	51,5	49,4	170,6	5,4
masculin	41,5	95,4	41,4	181,3	7,4
total	37,2	74	45,3	176	6,4
p	<0,001	<0,001	0,081		
B. Type de segment					
voyelle orale	20,8	100,7	19,6	169,7	3,1
voyelle nasale	45,2	79,6	38,8	220,6	9,9
consonne orale	12,6	100,3	30,2	165,9	2
consonne nasale	69	20,2	89,6	160,1	11,1
total	37,2	74	45,3	176	6,4
p	<0,001	<0,001	<0,001		
C. Contexte (pour V)					
CV	2,6	135	3,3	180	0,4
NV	39,1	77,6	35,9	169,3	6,3
VN	7,9	104,4	5,4	167,7	1,4
NVN	49,1	58,1	48,2	152,2	6,7
total	20,8	100,7	19,6	169,7	3,1
p	<0,001	<0,001	<0,001		
D. Type de V					
V avec corresp. Ṽ	14,6	112,6	13,5	167,4	2,1
V sans corresp. Ṽ	35,3	72,8	34	174,9	5,6
total	20,8	100,7	19,6	169,7	3,1
p	<0,001	<0,001	<0,001		
E. Type de C					
C occlusives	8,2	31,6	54,3	155,8	1,3
C fricatives	16,8	164,7	12,6	183,3	2,8
r et l	13,7	119,5	13,1	152,6	1,8
total	12,6	100,3	30,2	165,9	2
p	<0,001	<0,001	0,05		
F. Type de C (2)					
C sourdes	10,2	126,1	33,2	169,7	1,8
C sonores	14,3	67,8	35,1	168,4	2,2
r et l	13,7	119,5	13,1	152,6	1,8
total	12,6	100,3	30,2	165,9	2
p	0,008	<0,001	0,614		
G. Type de Ṽ					
ɑ̃	45,4	77,5	41,4	225,2	10,2
ɛ̃	26,5	101,2	21,6	232,4	6,2
ɔ̃	70	55,4	59	211,7	14,8
œ̃	29,1	91,2	25,1	199,8	5,7
total	45,2	79,6	38,8	220,6	9,9
p	<0,001	<0,001	<0,001		
H. Contexte (pour Ṽ)					
Ṽ	39,7	60,3	43,2	261,7	10,4
CṼ	44,3	81,2	37,9	227,9	9,9
NṼ	59,2	51,9	55,2	217,4	12,3
ṼN	33,2	87,7	29,5	124,2	4
CṼ.CV	42,5	106,8	30,2	147,1	6
CṼC	50	87,8	37,7	279,8	13,9
total	45,2	79,6	38,8	220,6	9,9
p	<0,001	<0,001	<0,001		

La nature de Ṽ. Les différences observées entre les quatre voyelles nasales sont significatives, ainsi que l'indique la table 2 (G). Seule la paire [ɛ̃-œ̃] ne connait pas de différence significative d'après les tests post hoc, ce qui ne manque pas d'intérêt, étant donné le statut précaire de cette opposition phonologique aujourd'hui en français. Par ailleurs, les données montrent qu'il existe une relation entre DAB et DAN pour les voyelles nasales: ces deux mesures sont (en première approximation) inversément proportionelles pour chacune des quatre Ṽ. Les cas les plus saillants sont ceux de [ɛ̃,œ̃] d'une part, et de [ɔ̃] d'autre part. A ouverture vélique et pression sous-glottique équivalente, la configuration radicalement différente du conduit vocal dans ces deux situations suffit à expliquer les valeurs obtenues : l'articulation postérieure et fermée de ɔ̃ induit une constriction entre le dos de la langue et le voile, qui a pour conséquence une plus grande résistance au passage de l'air et donc une réduction du DAB ainsi qu'une augmentation concomitante du DAN. Pour [ɛ̃,œ̃] c'est le contraire : ces voyelles sont antérieures et relativement ouvertes. Quant à [ɑ̃], c'est la plus ouverte des nasales du français, mais elle est aussi postérieure, ce qui génère une constriction relativement importante, au niveau de la paroi pharyngale cette fois.
Enfin, pour [ɑ̃] du moins – mais la question mérite d'être posée pour [ɔ̃], étant donné son taux de nasalité élevé – la nécessité d'un plus grand degré de nasalité à des fins de salience perceptuelle peut être invoquée à la suite des phénomènes observés dans [Hou56] et [Mae82].
Le contexte. De même que pour les orales, le taux de nasalité des voyelles nasales dépend du contexte dans lequel évolue la voyelle. L'ANOVA révèle que cette variable est significative (pour le DAN, pour le DAB et le PNA), et des tests post hoc ont précisé que les divers contextes étudiés peuvent se répartir en trois groupes : une voyelle nasale suivie de N a significativement moins de PNA que dans tout autre contexte, tandis qu'une voyelle nasale précédée de N en a (presque pour toutes les paires) significativement plus. Au milieu de l'échelle, on a les contextes Ṽ, CṼ, CṼC et CṼ.CV. Ceci indique encore une fois qu'une consonne nasale qui précède une voyelle favorise davantage sa nasalité qu'une consonne nasale subséquente. On peut même supposer un retard de l'abaissement du voile dans ce dernier contexte, étant donné que la nasalité observée pour Ṽ y est plus faible que lorsque celle-ci est, par exemple, précédée d'une consonne orale.

4. Discussion générale et conclusion

Deux phénomènes traversent l'ensemble des résultats analysés ci-dessus, l'influence générale du contexte sur la nasalité d'un segment et le réseau complexe de relations qu'entretiennent les mesures de DAB, de DAN et de PNA. Pour les voyelles (orales ou nasales), le fait d'être précédées par un segment consonantique nasal favorise leur taux de nasalité propre, alors que celui-ci est réduit si elles sont suivies d'une nasale. Les consonnes orales, par contre, ont un faible degré de nasalité, qu'elles soient suivies de Ṽ ou de V. Notre corpus n'ayant pas été constitué avec les consonnes orales comme préoccupation centrale, il ne présente que peu d'occurences de C précédées de Ṽ ou de V. Pourtant, l'examen de ces données invite à envisager d'une façon plus globale les résultats que nous venons de rappeler. En effet, les consonnes orales en contexte ṼC ont un DAN moyen très important (41ml/sec).
La table 3 représente schématiquement la réalisation en français de séquence de segments à valeur opposée pour le trait de nasalité, telle que suggérée par nos résultats. L'implémentation phonétique des deux types de séquence est tout à fait différente : à moins d'être bloquée, par exemple par une voyelle orale qui a une correspondante phonologique nasale, la nasalité déjà engagée a tendance à persister dans le segment suivant, alors qu'elle ne débute que rarement plus tôt qu'il n'est nécessaire. La commande phonologique conserve donc la prééminence en ce qui concerne le déclenchement de l'activité vélique, tandis que la nasalisation d'ordre phonétique intervient dans le maintien de cette activité.

Table 3: Taux de nasalité de séquences [+nasal] [-nasal] et [-nasal] [+nasal] en français

	[+nasal] [-nasal]		[-nasal] [+nasal]	
CV	++	+	0	+
VC	+	+	0	++

Pour conclure, nous insisterons sur le fait que les données que nous avons présentées ont mis en évidence l'étroite relation qu'entretiennent les valeurs de DAB et de DAN pour un type de segment donné. Etudier le DAN seul ne permet pas de mener une analyse fine de la nasalité. Les phénomènes aérodynamiques ne prennent sens que s'ils sont envisagés dans leur globalité. En particulier, les conditions de production entourant l'aspect "oral" d'un segment étudié pour sa nasalité influencent fortement ce second aspect "nasal", sans parler des variations de pression sous-glottique, dont il n'a pas été question ici. Enfin, les mesures quantitatives dont nous avons discuté dans cet article semblent à même de fournir une information fiable et détaillée sur ces phénomènes. Il s'agira à l'avenir d'analyser cette information concernant la production de la nasalité à la lumière des faits de perception, qui doivent notamment déterminer à partir de quand et dans quelle mesure les variations des paramètres aérodynamiques observées ici influencent la façon dont le message est décodé par l'auditeur.

Cette recherche est subventionnée par la Convention ARC 98-02 n°226

Bibliographie

[Tes90] Teston B., Galindo B. (1990) "Physiologia : un logiciel d'analyse des paramètres physiologiques de la parole", Travaux de l'Institut de Phonétique d'Aix, 13, pp.197-217.

[Huf93] Huffman M., Krakow R. (1993), "Instruments and techniques for investigating nasalization and velopharyngeal function in the laboratory : an introduction", Phonetics and phonology. Vol 5, pp.3-59.

[Hou56] House A., Stevens K. (1956), "Analog studies of the nasalization of vowels", Journal of Speech and Hearing Disorders, 21, pp.218-232.

[Mae82] Maeda S., "Acoustic correlates of vowel nasalization : A simulation study". JASA, 72, S102.

[Kin94] Kingston J., Diehl R. L. (1994). "Phonetic Knowledge", Language, 70, 3, pp.419-453.

Prosodie

Variations temporelles communiquant l'émotion dans la parole

S.J.L. Mozziconacci et D.J. Hermes*

IPO (Center for user-system interaction), Eindhoven, Pays-Bas
*Aussi : Institut de Phonétique d'Amsterdam, Amsterdam, Pays Bas
Lab. de Phonétique de l'Université de Leiden, Leiden, Pays-Bas
Mozziconacci@hotmail.com - http://www.fon.hum.uva.nl/mozzic

ABSTRACT

The present study investigates temporal characteristics of emotional speech. First, a production study was conducted at the global level of the whole utterance. Per emotion, the mean "global speech rate" was compared to the one found optimal in a previous perception study. Second, as more local information inside utterances might be specific to particular emotions, the study went a step further with the analysis of "local speech rate" relative to neutrality, considering accented and non-accented speech segments separately. The perceptual relevance of variations in local speech rate relative to neutrality was also tested in a perception experiment. Local variation appeared to be relevant for generating some emotions in speech while a linear manipulation appeared sufficient for other emotions.

1. INTRODUCTION

Divers types de variations telles que les variations mélodiques, d'intensité, de qualité de voix, de débit et de rythme de parole contribuent à l'expression et à la perception de l'émotion dans la parole. Divers travaux portent sur ces variations prosodiques, mais le nombre d'études quantitatives consacrées à l'aspect temporel de la parole exprimant de l'émotion [par ex. Bez84, Kit92] est limité, et les efforts se concentrent généralement sur des différences de débit global d'une émotion à l'autre. Le but de la présente étude est de cerner la contribution des variations temporelles à l'expression et la perception de l'émotion dans la parole en néerlandais. Les caractéristiques temporelles inter-émotions et inter-locuteurs sont décrites aussi bien au niveau global de la phrase entière qu'au niveau local de segments accentués et non-accentués dans des phrases exprimant une émotion. Cette étude de production est complétée par une étude de perception testant la pertinence perceptive des résultats.

2. DÉBIT AU NIVEAU GLOBAL

Le matériel utilisé consiste en 315 phrases résultant de l'enregistrement de trois locuteurs néerlandais (deux hommes, MR et RS, et une femme, LO) disant chacun trois fois cinq textes de contenu considéré sémantiquement neutre en exprimant les sept émotions : neutralité (en tant qu' «émotion» de référence), joie, ennui, colère, tristesse, peur et indignation. Afin de susciter ces émotions, les locuteurs ont d'abord dit des textes sémantiquement porteurs de l'émotion en question. La durée totale de chaque phrase a été mesurée. Les cinq textes n'induisant pas la réalisation de pauses, la notion de «débit global» peut être ici considérée de la façon la plus simple, comme inversement proportionnelle à la durée totale de la phrase. Cette notion de débit global est exprimée en tant que la proportion existant entre la durée moyenne des phrases exprimant une émotion et la durée moyenne des phrases sur le même texte dites en parole neutre par le même locuteur. Ainsi, par exemple, la valeur de débit '0.80' correspond à une réduction de débit résultant en un allongement de 20% de la durée de la parole émotionnelle par rapport à la parole neutre du même locuteur. Les résultats sont présentés dans la Table 1, séparément pour chaque locuteur. Le débit global moyenné sur les 3 locuteurs y figure aussi. Enfin, pour faciliter la comparaison des résultats de cette analyse à ceux de l'expérience de perception conduite au préalable [Moz98], les valeurs considérées optimales pour la perception figurent dans la colonne de droite de la Table 1. Dans cette étude de perception, des phrases manipulées en mélodie et en durée ont d'abord été présentées à des sujets devant indiquer par ordre de préférence les variantes leur semblant exprimer le mieux une émotion donnée. L'identification de ces émotions a ensuite été vérifiée en utilisant ces valeurs dans de la parole resynthétisée et de la parole synthétique.

Les résultats obtenus montrent que les 3 locuteurs sont souvent en accord sur le type de variation du débit global de parole. Ils sont unanimes pour exprimer la joie en utilisant un débit très proche de celui qu'ils adoptent en parole neutre, et pour ralentir leur parole en exprimant la tristesse et l'ennui. Par contre, lors de l'expression de la colère, alors que RS et LO ralentissent leur débit, MR l'accélère ce qui est

Table 1 : Débit global par locuteur, débit global moyen et valeurs trouvées optimales en perception.

Emotion	MR	RS	LO	Débit global moyen	Valeurs de perception
Neutralité	1.00	1.00	1.00	1.00	1.00
Joie	0.99	1.00	1.03	1.01	1.20
Ennui	0.73	0.87	0.85	0.82	0.67
Colère	1.16	0.84	0.82	0.94	1.27
Tristesse	0.93	0.87	0.96	0.92	0.78
Peur	1.15	1.25	0.93	1.11	1.12
Indignation	0.86	1.00	0.68	0.85	0.85

d'ailleurs consideré optimal suivant les resultats des expériences de perception. Pour la peur et l'indignation, c'est LO qui diffère des autres locuteurs en parlant relativement plus lentement qu'eux. Ces quelques différences témoignent de diverses stratégies d'expression pour une même émotion.

Les valeurs trouvées optimales à l'issue d'expériences de perception sont à plusieurs reprises plus extrêmes que les valeurs trouvées dans la parole produite par les locuteurs; là où les locuteurs réalisent une réduction, la valeur optimale suggère une réduction plus importante. Ceci est probablement dû au fait que lors des expériences de perception, seul le débit global était varié. En effet, pour s'exprimer en parole naturelle, on fait appel à divers paramètres et leurs effets combinés. Il semble raisonnable que les sujets, ne pouvant se fier qu'au débit pour identifier l'émotion, aient préféré des valeurs un peu plus extrêmes qu'en parole naturelle.

3. ANALYSE DE DÉBIT AU NIVEAU LOCAL

Sachant que les variations en débit moyen de parole contribuent à l'expression de l'émotion dans la parole, il semble souhaitable de considérer aussi la façon dont les variations temporelles sont réalisées à l'intérieur des phrases. Il se peut que l'expression de l'émotion exprimée détermine l'organisation temporelle au niveau local. Diverses options se présentent quant au choix des unités à considérer en fonction de l'émotion exprimée. Par exemple, l'allongement en position finale, la proportion de segments de silence dans la phrase, la façon dont les phonèmes sont affectés (par des réduc-tions, des assimilations, etc.), la durée des différents phonèmes étaient des candidats à l'étude. Néanmoins, la solution la plus simple, permettant d'étudier les variations temporelles à l'intérieur des phrases sans pour autant entraîner la prise en compte de trop de détails, est de considérer la durée relative des segments accentués et non-accentués. Un segment de parole accentué est composé d'une syllabe stressée lexicale-ment et sur laquelle un accent mélodique est réalisé. Un segment non-accentué est composé d'une syllabe ou de plusieurs syllabes consécutives sur lesquelles aucun accent n'est réalisé. Pour aller au plus simple et exclure aussi l'effet de l'allongement final, la dernière syllabe des phrases est simplement exclue des analyses.

Afin de distinguer l'effet de l'expression d'émotion de celui d'un simple changement de débit de parole, il est nécessaire de pouvoir comparer la durée de segments de parole émotionnelle à celle de segments de parole neutre produite à des débits variables. Une référence fiable n'ayant pas pu être trouvée pour le néerlandais dans la littérature, une analyse de parole neutre a été effectuée afin d'obtenir une référence. Le but est de déterminer si la proportion de segments de parole accentués et de segments non-accentués est similaire en parole émotionnelle et en parole neutre. Si c'est le cas, un modèle linéaire sera suffisant pour décrire les phénomènes temporels pertinents à l'expression de l'émotion dans la parole. Autrement, les règles qui régissent ces variations locales devront être déterminées, et la pertinence de variations locales sera testée en ce qui concerne la perception de l'émotion.

Parole émotionnelle Parmi les textes utilisés pour l'analyse globale précédente, deux n'ont pas été considérés adéquats pour la présente analyse, l'un parce qu'il induit la réalisation d'un accent en position finale, ce qui résulte en une interférence de l'allonge-ment final sur l'accentuation, l'autre parce que la présence d'une syllabe stressée supplémentaire a fait en sorte que les locuteurs n'ont pas systématiquement réalisé un accent unique sur la même syllabe dans ce texte. Chacun des trois autres textes contient deux syllabes stressées. Ci-dessous, les syllabes stressées sont soulignées, les segments accentués et non-accentués sont séparés par des traits verticaux, et les syllabes finales, exclues des analyses figurent en gris. Les textes sont les suivants : texte 1 : 'Zijn vrien | din | kwam met het | vlieg | tuig' (Son amie est venue en avion), texte 2 : 'Jan | is naar de | ka | pper ge | weest' (Jean est allé chez le coiffeur), et texte 3 : 'Het is | bij | na | ne | gen | uur' (Il est presque neuf heures). Au total, 182 des 189 phrases exprimant l'une des sept émotions (27 de ces phrases expriment la neutralité) ont été analysées. Sept phrases n'ont pu être considé-rées, soit que le locuteur ait bafouillé, soit qu'un allongement ait été produit à une frontière prosodique, soit que deux accents n'aient pas été réalisés dans la phrase. Comme ces sept phrases étaient des réalisations des trois différents locuteurs, exprimant cinq différen-tes émotions, il semble raisonnable d'assumer que le fait d'éliminer ces phrases n'influence pas les résultats.

Parole neutre Les trois mêmes textes ont fait l'objet d'un nouvel enregistrement du locuteur masculin MR, produisant, cette fois, de la parole neutre à des débits augmentant progressivement. L'étendue de ces variations en débit couvre celle des valeurs observées en parole émotionnelle. Au total, 171 phrases ont été enregistrées, 57 par texte.

Procédure La durée des segments accentués et non-accentués a été mesurée, du début du segment au début du segment suivant, dans les phrases chargées d'émotion et les phrases neutres à débit variable. Les durées des segments accentués et des segments non-accentués (à l'exclusion de la syllabe finale) ont été additionnées séparément. La proportion de la durée des segments accentués par rapport à la durée totale des segments a été calculée. Cette proportion sera plus brièvement nommée «proportion accentuée». Afin de constituer la référence de parole neutre, la ligne de régression optimale a été calculée pour chaque texte, sur la base des données de proportion accentuée pour la parole neutre. Les données obtenues dans la parole exprimant les émotions sont considérées par rapport aux données obtenues dans la parole neutre.

Résultats et discussion Les fonctions ci-dessous décrivent les lignes de régression optimales pour les données obtenues dans les phrases dites en parole neutre à débit variable. Pour les textes 2 («kapper») et 3 («uur»), le fait de considérer deux lignes de régression au lieu d'une seule augmente clairement la variation expliquée. Les fonctions suivantes, où x est la durée totale des segments en secondes, et y la proportion accentuée, sont représentées dans la figure 1.

Texte 1 : Pour tout x, $y = 0.4206 - (0.0388 \times x)$
Texte 2 : Pour tout $x < 1.83$, $y = 0.3246 + (0.0387 \times x)$
autrement, $y = 0.4845 - (0.0487 \times x)$
Texte 3 : Pour tout $x < 1.09$, $y = 0.3012 + (0.1026 \times x)$
autrement, $y = 0.4906 - (0.0712 \times x)$

Les résultats concernant les proportions accentuées dans la parole exprimant les émotions sont aussi présentés dans la figure 1, séparément pour chaque texte. Les points situés au-dessus des lignes de régression indiquent un allongement des syllabes accentuées plus important qu'en parole neutre. Le cas se présente par exemple pour l'expression de la colère, la tristesse, la peur et l'indignation sur le texte 1 («vliegtuig »). Il faut noter une limitation de la présente approche qui est due au fait que la référence de parole neutre à débit variable est basée sur la parole d'un seul locuteur, MR. De plus, sur le texte 3 («uur»), ce locuteur a produit différentes proportions accentuées dans la parole neutre à débit variable, et celle enregistrée exprimant l'«émotion de référence», la neutralité. Cette différence semble aussi se répercuter sur l'expression des émotions sur ce même texte.

Bien que l'analyse montre des différences de proportions accentuées et que, pour un débit global réduit, on perçoive une certaine tendance à allonger, en parole émotionnelle, les segments accentués relativement plus que les segments non-accentués, les résultats ne permettent pas la formulation d'hypothèses concernant la réalisation exacte des variations locales temporelles pour l'expression d'émotions spécifiques. Dans la suite de cette étude, on a recours à la perception pour tenter d'obtenir davantage d'informations sur la répartition des segments de parole exprimant les sept émotions.

3. IDENTIFICATION DE L'ÉMOTION EN FONCTION DU DÉBIT AU NIVEAU LOCAL

L'étude de production décrite précédemment montrant des variations locales sans pour autant être concluante en ce qui concerne l'influence de l'expression d'émotion sur la répartition des durées de segments de parole accentués et non-accentués, la présente étude se propose de tester si les variations locales sont pertinentes à la perception de l'émotion, et, si c'est le cas, de définir les valeurs optimales de proportion accentuée pour exprimer une émotion particulière.

Matériel Trois phrases neutres émises par le locuteur MR sur chacun des trois textes utilisés précédemment ont été manipulées par analyse-resynthèse. Deux séries de stimuli ont été générées, chacune constituée des six mêmes conditions. Dans la série 1, la durée totale des stimuli est constante, dans la série 2, cette durée varie suivant les valeurs trouvées optimales pour chacune des sept émotions dans une étude préalable [Moz98] (voir table 1). Les valeurs de «pitch level» et de «pitch range» trouvées optimales dans cette étude préalable ont été utilisées. La condition 1 ne présente aucune manipulation temporelle au niveau local. La proportion accentuée est variée dans les 5 autres conditions. Pour la condition 2, la proportion accentuée dépend entièrement du débit global, et est manipulée suivant

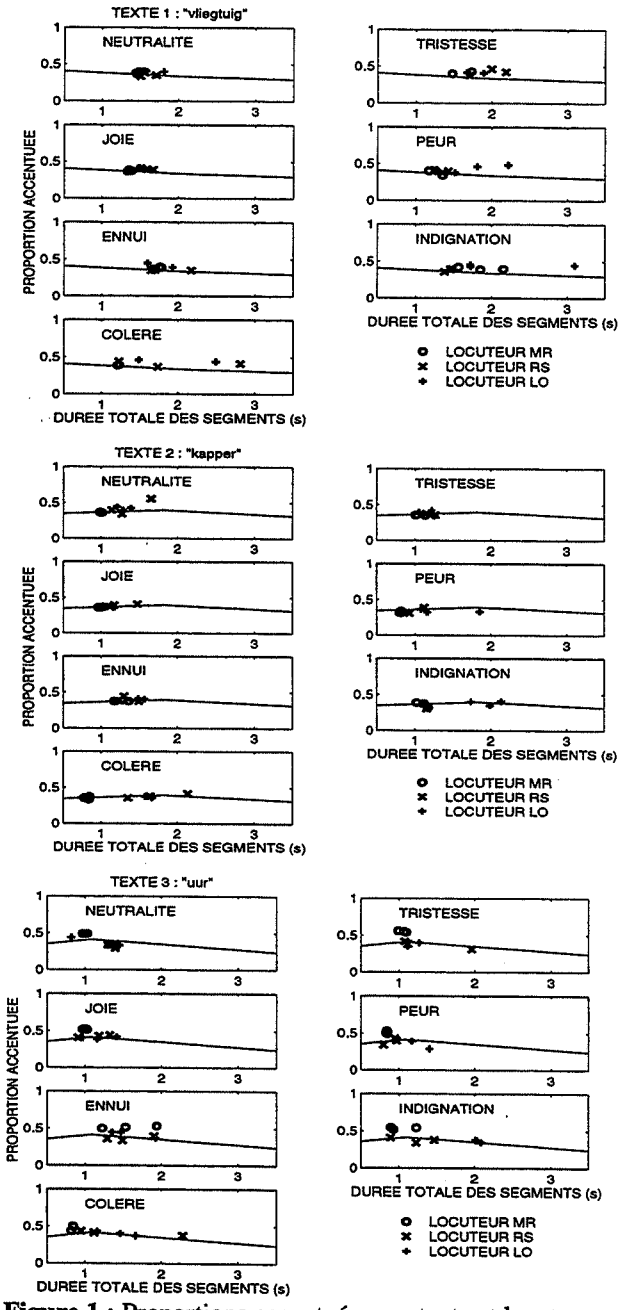


Figure 1 : Proportions accentuées par texte et locuteur

les fonctions décrivant les lignes de régression représentées figure 1 et issues de l'analyse de parole neutre à débit variable. Dans les conditions 3, 4 et 5, la

proportion accentuée est manipulée de façon à être respectivement réduite de 20%, augmentée de 20% et de 40% par rapport à la proportion accentuée «de référence» utilisée pour la condition 2. Tous les stimuli des conditions 1 à 5 ont été générés avec la configu-ration intonative '1&A 1&A' considérée acceptable dans l'expression de toutes ces émotions. Comme l'allongement de la voyelle a pour effet de rendre la descente mélodique 'A' plus clairement audible, ce qui pourrait influencer l'identification d'émotion, une sixième condition dans laquelle les stimuli sont générés avec la configuration '1B 1B' a été incluse afin de distinguer l'effet d'allongement de la syllabe de celui de l'audibilité du mouvement mélodique. Dans toutes les conditions, la syllabe finale n'a pas été affectée par les manipulations. Toutes les manipulations ont été basées sur la technique PSOLA [Mou95]. Au total, 252 stimuli (2 séries, 6 conditions, 3 textes, 7 combinaisons de pitch level et pitch range) ont été produits.

Procédure 24 sujets ont participé à ce test dans lequel les stimuli leur étaient présentés en 3 blocs (un par texte) dans un ordre aléatoire différent pour chaque sujet. Après avoir écouté la phrase une fois, ils devaient choisir, parmi les 7 catégories d'émotions proposées, celle qui selon eux avait été exprimée.

Résultats Les résultats des tests d'identification ont subi une analyse log-linéaire [Fie80]. Le modèle décrivant le mieux les données est celui où il existe des effets significatifs de PITCH (les combinaisons de «pitch level» et «pitch range»), COND (les conditions), RESP (les réponses des sujets), et des interactions significatives entre PITCH et RESP, et entre COND et RESP. Une analyse de clusters a permis de former deux clusters de chacun trois conditions. Le premier cluster réunit les trois premières conditions, correspondant à de faibles proportions accentuées. Les trois dernières conditions forment le cluster des assez grandes proportions accentuées. Le fait que les conditions 5 et 6 fassent partie du même cluster indique que l'effet qui influe sur les réponses des sujets est lié à la longueur des syllabes et non à l'audibilité du mouvement mélodique. Les résultats sont présentés dans la table 2 pour les deux clusters ainsi formés. Les flèches indiquent que le nombre de réponses des sujets dans la catégorie correspondante est, de façon significative, plus (↑) élevé ou moins (↓) élevé que selon les prédictions d'un modèle log-linéaire représentant l'absence d'effet des conditions sur ces réponses. Des déviations significatives de ce modèle ne sont obtenues que pour les catégories neutralité et indignation, mais dans ces deux cas, ces variations temporelles s'avèrent très importantes pour la perception. La proportion accentuée doit être limitée en parole neutre. Une augmentation de la proportion accentuée se fait au détriment de la perception de neutralité. En l'occurrence, une proportion accentuée qui est importante suggère la perception de l'indignation.

Table 2 : Nombre de réponses par cluster de conditions

Conditions	Réponses des sujets						
	Neutr	Joie	Ennui	Colère	Trist	Peur	Indign
Série 1 Cond. 1 à 3	566↑	290	63	98	187	146	162↓
Cond. 4 à 6	387↓	259	102	92	206	159	307↑
Série 2 Cond. 1 à 3	280↑	236	328	179	184	136	169↓
Cond. 4 à 6	169↓	223	338	174	160	159	289↑

4. Conclusion

Il a été confirmé que les variations temporelles au niveau global de la phrase entière sont d'une importance primordiale, tout particulièrement pour certaines émotions comme l'ennui. En comparaison à cet effet majeur de débit global ou celui de la mélodie, les variations en proportion accentuée ont relativement peu d'effet, ce qui n'a rien d'étonnant. Bien que l'analyse de production n'ait montré qu'une tendance à varier la structure temporelle lors de l'expression d'émotion, le test de perception a démontré que les variations en proportion accentuée remplissent une fonction communicative, telle que celle de signaler certaines émotions dans la parole.

Remerciement

Ce travail bénéficie du soutien du projet CREST, JST (Japan Science & Technology), Kyoto, Japon.

Bibliographie

[Bez84] Bezooijen R.A.M.G. van (1984), The characteristics and the recognizability of vocal expression of emotion, Foris, Dordrecht, The Netherlands.

[Fie80] Fienberg S.E. (1980), The analysis of cross-classified categorical data, MIT Press, Cambridge, Massachusetts.

[Kit92] Kitahara Y. and Tohkura, Y. (1992), "Prosodic control to express emotions for man-machine interaction", IEICE Transactions on Fundamentals of Electronics, communication and computer sciences, 75, 155-163.

[Mou95] Moulines E. et Laroche J. (1995), "Non-parametric techniques for pitch scale and time-scale modification of speech," Speech Communication, 16, 175-205.

[Moz98] Mozziconacci S.J.L. (1998), Speech variability and emotion : production and perception, Eindhoven.

L'implication emphatique dans la narration orale spontanée : validation perceptive et réalisations acoustiques

Odile Bagou°, Albert Di Cristo°*

° Laboratoire Parole et langage, Université d'Aix en Provence
*Université de Genève, FAPSE, Laboratoire de Psycholinguistique expérimentale
40 boulevard du pont d'Arve CH 1205 GENEVE
e-mail : odile.bagou@pse.unige.ch Albert.DiCristo@lpl.univ-aix.fr

ABSTRACT

The present paper shows on one hand that two typical emphasis categories may be distinguished: the lexical emphasis (EL) and the supralexical emphasis (ESL), depending on the domain of linguistic achievement. On the other hand, we state that judges can identify not only the domain extent but also the degree of implication in speech productions. The ESL group is percieved as having greater emphasis than that of the EL group. Our corpus, recorded by male speakers, allows us to demonstrate that emphasis is generally marked by widening pitch range and increasing f0 peaks on those syllables bearing emphasis, with a significant implication apogee. Moreover, we have also found that an increase in emphasis degree goes hand in hand with an increase in pitch range and more important f0 peaks on the emphasis bearing syllables. Whilst ESL and EL are percieved differently, we failed to find the relevant acoustical cues.

1. INTRODUCTION

Les descriptions de l'emphase relatées dans la littérature, présentent un flou théorique considérable [Bag98, pour une revue] se résumant, d'une part, à une limitation de cette notion et d'autre part, à la diversité des domaines de réalisation linguistique étudiés. En effet, dans le cadre d'analyses phonétiques, ce phénomène est généralement considéré comme un « accent d'insistance », ce qui limite l'analyse à un traitement de l'information acoustique localisée. De plus, peu d'analyses instrumentales ont été menées sur ce phénomène. Cependant, les travaux sont effectués au niveau de la syllabe [Seg77] [Tou87], voire au niveau du mot [Dah96], mais aucune étude n'a, à notre connaissance, abordé l'emphase sur une unité supérieure au mot.

A l'opposé, l'approche linguistique tend à apprécier l'emphase comme une marque d'implication du locuteur dans son discours, sans en définir exhaustivement les réalisations acoustiques.

Une autre perspective de recherche a été d'associer l'emphase à un accent « émotionnel », celle-ci négligeant alors son aspect linguistique. En effet, certaines études ont tenté d'établir des corrélations entre les paramètres acoustico-prosodiques et les attitudes du locuteur par rapport à son discours, [Cou86] [Cry69]. Cependant, certains de ces travaux sont basés sur l'introspection et utilisent des notions préétablies par l'expérimentateur [Bol86].

Enfin, certaines études pragmatico-acoustiques, nous semblent mieux définir le phénomène. Cependant, dans tous ces travaux, l'analyse est réalisée dans un unique type d'interaction, la conversation, ce qui ne permet, en aucun cas, une généralisation du phénomène d'emphase [Fie90] [Sel94].

La démarche théorique que nous proposons d'adopter dans ce travail, est une approche multidimensionnelle. En effet, nous considérons l'emphase, à la fois comme étant la résultante d'une implication du locuteur dans son discours, tout en ne négligeant ni sa dimension émotionnelle, ni sa capacité à apporter des réponses localement pertinentes [Spe89], en adéquation avec la situation et le contexte de production. En effet, l'interprétation du langage étant relative au contexte, des réalisations marquées telles que l'emphase, peuvent permettre, d'une part au locuteur de rendre certaines unités saillantes par rapport aux usages environnants, et d'autre part à l'allocutaire de réduire les éventuelles hypothèses émises quant à l'intention profonde du locuteur. Afin de considérer la nature multidimensionnelle de notre étude, nous définissons également l'emphase en fonction de ses corrélats acoustico-prosodiques.

Enfin, nous travaillons sur la narration orale spontanée, ce qui nous permettra de savoir, dans de futurs travaux et par comparaison avec les études effectuées dans la conversation, si l'emphase est caractérisée différemment, d'un style de parole à un autre.

D'autre part, les divergences théoriques dont nous avons fait préalablement état, induisent une diversité des domaines de réalisation étudiés. L'approche phonétique assimile généralement l'emphase à un accent local, ce qui limite l'étude à la syllabe.

Nous préférons étendre la compréhension du phénomène en ajoutant, dans ce travail, sa dimension linguistique. Nous montrerons que l'emphase ne se résume pas à une proéminence syllabique locale. En effet, il apparaît que, sur l'axe syntagmatique, ce phénomène se définit par des domaines de réalisation linguistique plus larges. En conséquence, nous définirons le(s) domaine(s) linguistique(s) de l'emphase, et chercherons leurs corrélats prosodiques. Enfin, après avoir mis en évidence ces diverses catégories linguistiques, nous montrerons que ce phénomène est continu sur l'axe paradigmatique, c'est à dire qu'il existe plusieurs degrés d'implication emphatique.

Dans la présente étude, nous tentons donc d'analyser les « règles affichées » [Cou86], c'est à dire les marques formelles d'une planification pragmatique sous-jacente utilisées par le locuteur à la fois pour manifester son implication croissante dans son discours aboutissant en un point culminant, le « climax » [Bol86], et pour « contextualiser » ses attitudes. Pour ce faire, nous procèderons à une validation perceptive et à une étude prosodique (gamme mélodique et pics de f0) des productions.
Avant d'entrer dans le détail de l'exposé de la méthode et des résultats obtenus, il nous semblait important de soulever le problème de la circularité éventuelle d'une telle approche. Afin de pallier ce problème crucial, Nous adoptons une démarche fonctionnelle se focalisant sur les mécanismes d'implication dans la construction de présupposés partagés plutôt que sur des listes de signes de l'implication. L'objectif général étant d'analyser ces marques formelles dans plusieurs types d'interactions et de situations de productions différentes, nous avons préalablement privilégié l'analyse de la narration orale spontanée, celle-ci attestant d'usages marqués plus nets.

2. METHODE

2.1. Procédure expérimentale

Notre population de juges-auditeurs est constituée de 10 sujets francophones, âgés de 20 à 53 ans (moyenne : 31.8 ; Ecart type : 12.1 ; Etendue : 33). Lors de la première partie du test, la tâche proposée était de souligner précisément, à partir de productions transcrites orthographiquement, les fragments de narration qu'ils percevaient comme la manifestation d'une « implication insistante » du locuteur, dans sa production langagière. Puis, ils devaient entourer la syllabe qui leur semblait correspondre à une apogée emphatique.
Dans un second temps, nous leur demandions d'attribuer un degré d'implication, situé sur une échelle de magnitude en 9 points pour chaque partie préalablement soulignée.
Les mesures effectuées à l'issu de ce test de validation perceptive, sont l'empan de l'emphase, c'est à dire son domaine de réalisation linguistique, et le degré d'implication.

2.2. Caractérisation acoustique des unités emphatisées

Nous avons recueilli les productions de narrations orales spontanées du 'petit chaperon rouge' en chambre sourde, afin de garantir une qualité d'enregistrement suffisante. Nous avons demandé à quatre locuteurs masculins de raconter cette histoire à un enfant, sans support textuel écrit. Les événements pertinents sont conservés sur la base du test de validation précédemment décrit. Seuls les phénomènes reconnus par au moins 70% des juges sont considérés comme méritant une analyse plus précise.
Après avoir étiqueté les signaux acoustiques en syllabes, à partir de critères de segmentation stricts [Bag98], nous avons procédé à une modélisation de la courbe de F0, afin de ramener celle-ci à quelques unités discrètes, des points cibles attestant de changements de direction phonologiquement pertinents (INTSINT).
Malgré la nature pluri-paramétrique de la prosodie, nous avons choisi de ne mesurer que les gammes mélodiques et les pics de fréquence fondamentale, aux dépends de la durée et de l'intensité. En effet, certains critères relevés dans la littérature, ont motivé le choix d'une analyse prioritaire de f0. De nombreux travaux mettent en évidence la pertinence de ce paramètre dans l'emphase et ce, dans plusieurs langues du monde. En français, la fréquence fondamentale est le paramètre privilégié de la réalisation du phénomène, dans 82% des cas. En revanche, l'intensité ne permet d'expliquer que 53% des cas d'emphase [Seg77]. Malgré une augmentation systématique de la durée démontrée dans divers travaux, le fait que la durée intrinsèque des syllabes ne soit pas systématiquement comparable, nous contraint à ne pas considérer ce paramètre. De plus, des études plus récentes s'attachant à démontrer la nature catégorielle de la perception de l'emphase, fondent également leurs hypothèses sur la supériorité de la fréquence fondamentale [Gus88].

3. RESULTATS

3.1. Analyse perceptive

Les résultats de l'analyse du test de validation nous permettent de dissocier, sur l'axe syntagmatique, deux types d'emphases définies par leur domaine de réalisation linguistique : les emphases lexicales (EL), portant sur le mot, et les emphases supra-lexicales (ESL), portant sur une unité supérieure au mot (figure1).

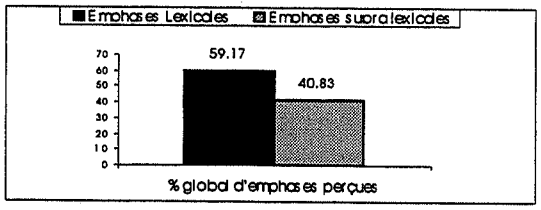


Figure 1 : les divers domaines de réalisation linguistique de l'emphase

D'autre part, le test de validation nous a permis d'extraire, à partir du jugement des auditeurs, 3 catégories d'emphases définies à partir du degré d'implication perçu. Il apparaît que l'implication du locuteur se manifeste préférentiellement par des emphases perçues comme moyennes ou fortes, c'est à dire de degré II ou III, plutôt que par des emphases de degré I (figure 2).

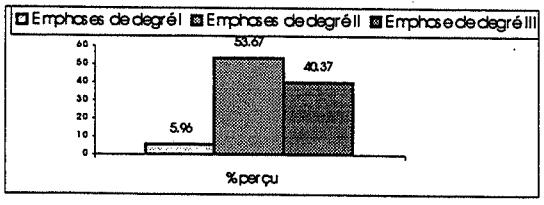


Figure 2 : Divers degrés de perception de l'emphase

Enfin, nous avons observé que la majorité des EL est considérée comme manifestant une implication moyenne, de degré II, alors que les ESL sont évaluées comme étant fortes, de degré III (figure 3).

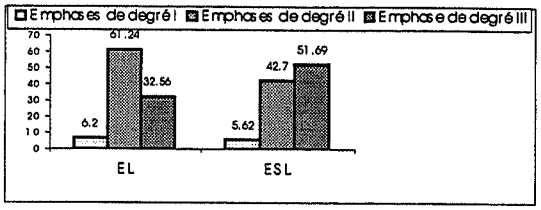


Figure 3 : analogie entre degré d'implication et empan

3.2. Analyse acoustique

Les résultats de l'analyse des phénomènes pertinents montrent que l'élargissement de la gamme mélodique, calculé en demi tons, est un paramètre acoustique caractéristique de la parole emphatique. En effet, 100% des productions marquées par une emphase présentent cette caractéristique, par rapport aux usages non marqués (figure 4).

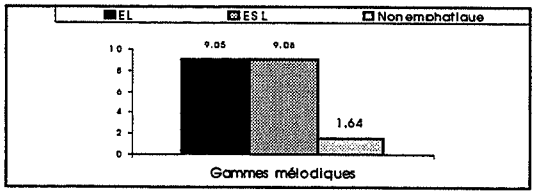


Figure 4 : présentation des gammes mélodiques de chaque catégorie d'emphase (EL/ESL) vs usages non marqués

Il apparaît que l'augmentation de la largeur de gamme mélodique est relative au degré d'implication perçu par les juges. En effet, les résultats montrent qu'un élargissement de la gamme mélodique induit la perception d'une implication plus forte du locuteur dans l'énoncé (figure 5).

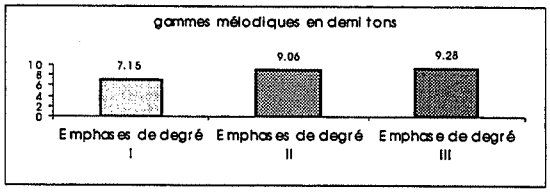


Figure 5: variations de la gamme mélodique en fonction du degré d'implication

En revanche, les résultats ne sont pas significatifs en ce qui concerne les différences observées entre EL et ESL, du point de vue de leurs largeurs de gammes mélodiques respectives. L'analyse instrumentale des pics d'implication montre d'une part, que les usages marqués sont caractérisés par une augmentation nette de la valeur de f0 sur les événements saillants par rapport aux usages environnants et que celle-ci varie proportionnellement au degré de perception de l'implication. (figure 6).

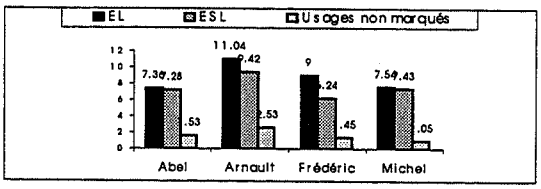


Figure 6 : présentation, par locuteur, des valeurs des pics de f0 en fonction du degré de perception de l'implication

D'autre part, il apparaît que les valeurs de f0 sur les pics d'implication sont plus faibles dans le cas des EL que dans le cas des ESL. Dans 60% des cas, les ESL attestent d'une fréquence fondamentale plus élevée sur les syllabes portant l'apogée (figure 7). Toutefois, ces données ne sont pas significatives et ne permettent pas de confirmer notre hypothèse de départ.

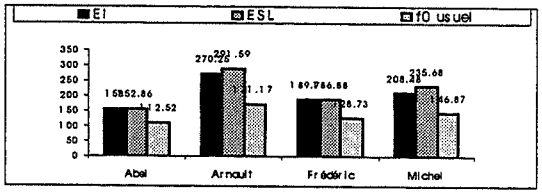


Figure 7 : présentation, par locuteur, des valeurs des pics de F0 en fonction du domaine linguistique d'implication

En résumé, nos résultats montrent que la fréquence fondamentale -et notamment la largeur de gamme mélodique et les pics d'implication- est un paramètre acoustique très significatif de l'emphase. Nous discuterons ces résultats ultérieurement.

4. DISCUSSION

Supposant que l'implication emphatique est planifiée par le locuteur en fonction de la situation de production et de son intention profonde, nous avons montré que la structure de la narration orale spontanée induisait des implications nombreuses et variées. En effet, il apparaît que les séquences narratives étudiées attestent d'une présence majoritaire d'implications moyennes ou fortes (de degré II ou III), et d'un taux négligeable d'implications faibles (degré I). Ceci corrobore notre postulat de départ selon lequel le récit oral induit une plus grande implication du locuteur que d'autres activités langagières [Goo84]. Cette étude préliminaire nous a permis de dissocier, à la fois sur l'axe syntagmatique et sur l'axe paradigmatique, plusieurs catégories d'emphases. D'une part, ce travail nous a permis de vérifier l'hypothèse selon laquelle l'implication peut se réaliser, sur l'axe syntagmatique, dans des domaines linguistiques supérieurs à la syllabe ou au mot. Nous avons mis en évidence deux types d'emphases, les EL et les ESL, c'est à dire les manifestations de l'implication planifiées respectivement sur le mot et sur une unité supérieure à celui-ci. D'autre part, Nous avons montré que diverses catégories pouvaient être mises en évidence sur l'axe paradigmatique. A partir du jugement des auditeurs, nous avons classé les emphases en trois grands

types : les emphases de degré I, II, et III, correspondant respectivement à des implications faibles, moyennes ou fortes du locuteur dans son dire. Enfin, supposant qu'une planification sur une unité linguistique large pouvait susciter une implication croissante, l'examen des jugements nous a permis de montrer que l'empan avait une incidence directe sur le degré d'implication perçu, les ESL attestant d' implications de degrés supérieurs.

L'objectif principal de l'analyse acoustico-prosodique à laquelle nous avons procédé était d'une part, de confirmer que la fréquence fondamentale est un paramètre pertinent dans la réalisation acoustique de l'emphase, et d'autre part, de mettre en évidence une corrélation entre ce paramètre et la perception des juges. Dans cette perspective, les ESL étant perçues comme plus emphatiques que les EL, il nous semblait cohérent de penser que celle-ci attesteraient de variations acoustiques plus importantes. A l'issu de cette analyse, nous pouvons conclure à une augmentation de la gamme mélodique dans les usages emphatiquement marqués par rapport aux usages non marqués. Toutefois, nous ne pouvons, en aucun cas, confirmer l'hypothèse d'une augmentation des valeurs de gammes mélodiques en fonction de la force d'implication perçue. En effet, les ESL sont considérées par les juges comme représentatives d'une implication plus forte du locuteur dans son discours, mais n'attestent pas pour autant de gammes mélodiques significativement plus larges que les EL. En revanche, nos résultats montrent que le degré d'implication perçu est corrélé de manière significative, à l'élargissement de la gamme mélodique. Une gamme mélodique étroite induit en effet, la perception d'une implication modérée, voire faible.

De la même manière, les résultats de l'analyse des mesures de f0 sur les syllabes portant l'apogée d'implication, montrent que les usages marqués sont caractérisés par une nette augmentation par rapport aux usages environnants non marqués. Il apparaît que le domaine de réalisation linguistique de l'implication ait une incidence sur les valeurs de f0. Toutefois, les résultats ne corroborent pas nos hypothèses, puisqu'il apparaît que les ESL sont perçues comme plus emphatiques que les EL, sans que les valeurs de fo sur les apogées d'implication ne soient significativement plus élevées pour les ESL que pour les EL. Nous pouvons expliquer ce résultat par le fait que la ponctualité des EL contraint le locuteur à planifier une augmentation plus rapide et plus forte de la hauteur mélodique, pour que son implication soit perçue par l'auditeur. Nous pouvons dire, à la lumière de ces résultats et conformément aux résultats du test de validation, que les ESL requièrent une planification sous jacente plus précoce qui permet au locuteur de signaler progressivement son implication, alors que les EL sont planifiées peu de temps avant la réalisation et nécessitent, par voie de fait, un effort plus grand, pour un effet moindre sur l'auditoire.

Comme dans le cas de l'élargissement de la gamme mélodique, les valeurs de f0 augmentent proportionnellement au degré d'implication perçu. Il nous semblait cohérent de penser que le sexe des locuteurs pouvait être la cause de tels résultats. Nous avons donc réalisé les mêmes mesures sur la base d'une production féminine. Il apparaît que la fréquence fondamentale varie de manière inverse, en ce sens que les valeurs de f0 décroissent parallèlement à l'augmentation du degré d'implication de la locutrice dans son discours. Cette remarque nous invite à suggérer l'existence de plusieurs stratégies quant aux variations acoustiques de la parole dans le cas d'usages marqués, mais nous ne pouvons aucunement conclure à l'existence de différences inhérentes au sexe du locuteur. Des travaux ultérieurs seraient nécessaires pour confirmer cette hypothèse.

Enfin, d'après ces quelques résultats, il nous semble intéressant de poursuivre l'analyse acoustique en considérant la nature pluri-paramétrique de la prosodie.

D'autre part, les études futures devraient permettre de vérifier l'hypothèse de l'existence d'une variabilité de planification et de réalisation liées au type d'activité langagière mobilisé. Nous tenterons donc de décrire d'éventuelles récurrences, attestées pour un même locuteur, dans des situations de productions diverses, afin d'extraire une information relative à l'existence de phonostyles caractéristiques. Cette étude comparative nous permettrait de vérifier des hypothèses relatives à l'émergence d'un style particulièrement emphatique dans certaines situations de productions, et de mettre en exergue des caractéristiques récurrentes de chaque type d'activités langagières.

REMERCIEMENTS à C. Fougeron pour ses précieux conseils de rédaction

6. BIBLIOGRAPHIE

[Bag98] Bagou, O. (1998) *L'implication emphatique*. Mémoire de DEA; Université d'Aix en Provence.

[Bol86] Bolinger, D. (1986), *Intonation and its parts*. Stanford, CA: Stanford University Press.

[Cou86] Couper-Khulen, E. (1986), *An introduction to english prosody*, Tübingen: Niemeyer.

[Cry69] Crystal, D., (1969), *Prosodic systems and intonation in English*. London: Cambridge University Press.

[Dah96] Dahan, D., Bernard, J.M., (1996), «Interpseakers variability in emphatic accent production in french », *language and speech*, 39; **4**: 341-374.

[Fie90] Fiehler, R. (1990) *Kommunkation und emotion*, Berlin: De Gruyter.

[Gus88] Gussenhoven, C. et Rietveld, T. (1988), « Fundamental frequency declination in dutch : testing three hypothesis », *Journal of phonetics*, **16**: 355-369.

[Sel94] Selting, M. (1994) « Emphatic speech style-with special focus on the prosodic signqlling of heihtened emotive involv,ent in conversation », Journal of pragmatics, **22**: 375-408.

[Seg77] Séguinot, A. (1977), "L'accent d'insistance en français standard" In *L'accent d'insistance: emphatic stress*, Eds F. Carton; D. Hirst; A. Marshall; and A. Séguinot, 12: 1-58. Paris: studia phonetica, Didier.

[Tou87] Touati, P. (1987), *Structures prosodiques du suédois et du français: profils temporels et configurations tonales*. Lund: Lund University Press.

Configurations prosodiques et thématisation dans la lecture à voix haute. Approche comparative.

Marie-Ange ALEXANDRE & Claire GERARD

Laboratoire Langage et Cognition, UMR CNRS Université de Poitiers
MSHS 99, avenue du Recteur Pineau, 86022 Poitiers Cedex, France
Tel : ++33(0)549 45 46 02
Mél: Marie-Ange. Alexandre@mshs.univ-poitiers.fr

ABSTRACT

Reading aloud texts which include thematized passages results in prosodic variations on the selected theme. Acoustic analyzes were performed on a total of 15 measures, global and local, including measures of voice intensity. In general terms, thematization results in three major types of modifications of speech in adults: A lower reading speed and a shift in the word accent. The analysis of fundamental frequency indicators, and of phrases and words intensity and duration shows several types of strategies, or speech styles, that are stable in adult speakers, and under construction in children.

Keywords: reading aloud, prosodic configurations, strategies and styles.

1. INTRODUCTION

1.1. Spécificité de l'étude.

Cette recherche de psychologie cognitive aborde le problème de la transmission des informations en référence à des représentations mentales construites par chaque individu. Elle vise à préciser les modifications de la parole qui interviennent pendant la thématisation, dans la lecture à voix haute de textes. Durée et fréquence fondamentale de la voix sont classiquement étudiées [Cut80, Sor89, Gér98]. Nous avons intégré, de plus, l'étude de l'évolution de l'intensité de la voix, ce qui est rarement réalisé dans les études sur la prosodie en raison de difficultés liées au recueil et au traitement des données. L'utilisation des mesures d'énergie développées dans notre laboratoire et les programmes mis au point pour les calculs de moyenne, écart-type, étendue et contour d'intensité ont résolu ces problèmes. Nous avons privilégié la recherche d'indices acoustiques globaux et locaux susceptibles de varier avec la thématisation réalisée, tout en étudiant à la fois ce que les locuteurs avaient en commun, et ce qui les différenciait, donc en précisant leur style. La variabilité interlocuteurs, mentionnée par de nombreux auteurs travaillant sur la parole, nous semble alors pouvoir être précisée, en dégageant ces ressemblances et différences.

1.2. Accent d'insistance et configurations prosodiques.

L'accent d'insistance est utilisé pour marquer certains mots de manière volontaire et se distingue donc de l'accent tonique (ou « de langue »). Par nature, il est placé par le locuteur même, en fonction de ses intentions communicatives, alors que l'accent tonique est fixé par les règles inhérentes à la langue française. Pour Fonagy [Fon79], la mise en relief d'une syllabe passe par un plus grand effort expiratoire et articulatoire, entraînant des modifications conjointes de vitesse d'articulation, de fréquence fondamentale et d'intensité de la voix. Pour Rossi [Ros79], l'accent d'insistance se distingue de l'accent tonique par trois facteurs principaux : il est initial alors que l'accent tonique est généralement final en français ; il est facultatif, contrairement à l'accent tonique qui, par définition, est obligatoire ; enfin, son domaine d'application est le mot ou la syllabe (lexical ou segmental), alors que celui de l'accent tonique est le groupe de mots ou la phrase (suprasegmental). Vaissière [Vai80] a introduit la notion de « groupe » ou de « mot » prosodique, unité principale du traitement prosodique qui peut se définir par son contour prosodique, c'est-à-dire par une configuration particulière de durée, d'intensité et de fréquence fondamentale. Lors de la production d'une séquence de parole, les contours prosodiques de chaque mot se combinent entre eux de façon cohérente et forment un patron de réalisation acoustique homogène. Un contour prosodique unique peut donc être superposé à plusieurs mots lexicaux. Vaissière a extrait quatre grands types de contours prosodiques du français, basés sur leur courbe particulière de fréquence fondamentale. L'étude présentée ici vise notamment à compléter cette approche par l'examen conjoint de mesures d'intensité, de fréquence fondamentale et de durée, et à développer la notion de « configurations prosodiques ».

Ces *configurations prosodiques* conditionneraient le niveau d'attention que des auditeurs allouent à l'estimation des variations des indices prosodiques [Sor89, Gér95]. Ainsi, sur la base de ses attentes (liées au contexte phrastique, conversationnel...), un auditeur émettrait une « hypothèse » quant à l'arrangement des paramètres prosodiques à venir. Puis, en comparant l'arrangement finalement perçu à un registre de configurations pré-stockées, il serait en mesure de détecter des contrastes fins de fréquence fondamentale, d'intensité et de durée. Ce registre de configurations

prosodiques apparaît donc comme un élément primordial des interactions verbales entre individus, se construisant chez l'enfant jusqu'à devenir opérationnel chez l'adulte.

1.3. Développement des configurations prosodiques.

Gérard et Clément [Gér98] et Clément et Gérard [Clé96], ont étudié la production et l'identification de différentes formes prosodiques chez des enfants de 5 à 9 ans. Dans des tâches de production d'énoncés, en utilisant une procédure de neutralisation sémantique, elles montrent que les compétences prosodiques sont acquises graduellement tout au long du développement. Et alors que les enfants contrôlent la hauteur de leur voix avant leur vitesse de parole en ce qui concerne l'expression de formes prosodiques relevant du mode linguistique (assertion, question), tous les paramètres sont déjà contrôlés dès l'âge de 5 ans, pour ce qui concerne l'expression du mode émotionnel (joie, tristesse). Nous savons que lors de la lecture à voix haute de phrases ou de textes, les locuteurs organisent spécifiquement leur prosodie pour transmettre un accent d'insistance [Gér95], et que les compétences semblent se construire progressivement tout au long du développement [Clé96]. Nous supposons que la thématisation, étape du processus de compréhension d'un texte, sera traduite à un niveau comportemental, non seulement par l'allongement des temps de lecture des parties cibles, mais aussi par une augmentation de la fréquence fondamentale et de l'intensité de la voix, et qu'elle ne sera pas opérationnalisée de la même manière en fonction du lecteur [Vai80]. Nous cherchons à observer la mise en pratique de stratégies différentes ou styles de lecture, qui reflètent selon nous, non seulement les capacités de conceptualisation, de programmation et de gestion de l'articulation mais aussi la maturation des processus mentaux impliqués dans ce type de tâche (décodage grapho-phonémique, processus de compréhension, de thématisation ...).

2. METHODE

Sujets : Trois femmes françaises, quatre enfants de 8 ans et six enfants de 10 ans ont participé à l'expérience.
Matériel : Des histoires courtes, développant chacune deux thématiques distinctes, avaient la même structure de base : une introduction qui présentait la situation, suivie de deux paragraphes (ou « sous-thèmes »). Les textes s'achevaient par quelques phrases de conclusion renvoyant au thème général de l'histoire. Chacune des histoires était proposée avec deux titres différents, renvoyant respectivement à l'un ou l'autre des sous-thèmes. La partie du texte qui devait être thématisée était encadrée. La consigne précisait le but de la lecture : transmettre le thème sélectionné dans le texte à un auditeur éventuel.
Procédure : Une étude à voix basse pendant quelques minutes (afin de se saisir du texte) précédait la lecture à voix haute. Les adultes lisaient 8 textes (les 4 histoires, chacune présentée deux fois pour les thèmes 1 et 2). Les enfants ne lisaient que six textes (3 histoires sur les 4, présentées deux fois).
Les enregistrements sonores ont été segmentés[1] en portions de phrases comprises entre deux signes de ponctuation – appelés plus bas « syntagmes » -, et en "mots-cibles". Ces mots-cibles ont été choisis, pour des raisons lexicales et sémantiques, comme susceptibles de porter l'accent d'insistance.
Les indices acoustiques[2] sont analysés à deux niveaux : 1- celui des syntagmes (niveau suprasegmental) et 2- celui des mots-clés (niveau lexical). Au niveau des syntagmes, nous avons mesuré la durée des pauses ponctuées et la durée d'énonciation des syntagmes, la fréquence fondamentale moyenne et l'écart-type des valeurs de fréquence fondamentale, l'intensité moyenne et l'écart-type des valeurs d'énergie. Au niveau des mots-cibles, nous avons mesuré la durée d'énonciation des mots, leur intensité moyenne, l'écart-type et la gamme de variation des valeurs d'énergie ainsi que le contour intensif ; la fréquence fondamentale fait l'objet des mêmes mesures (moyenne, écart-type, gamme et contour). Nous considérons que l'allongement des temps de pause et des temps de lecture, et une augmentation des valeurs moyennes d'intensité (mesurées en décibel) et/ ou de fréquence fondamentale (mesurées en Hertz) sur certaines parties peuvent représenter la thématisation opérée par le lecteur. Les écarts-types et les gammes de variation de ces deux derniers paramètres sont des indices de l'expressivité des lecteurs ; plus ils sont importants et moins le lecteur a une voix monotone. Deux indices, compris entre 0 et 1, opèrent sur le contour mélodique et le contour intensif. Le premier a été mis au point par Eady et Cooper [Ead86], il donne une indication quant à la position du pic de F0 au sein d'un mot[3]. Le deuxième, l'indice de contour intensif, témoigne de la place du pic d'énergie dans le mot. Nous supposons que les pics de F0 et d'énergie sont déplacés par l'insistance, vers le début des mots.

3. RESULTATS.

Nous avons étudié systématiquement l'effet du facteur « *Thématisation* », et l'*interaction* entre ce facteur et le facteur « Sujet ». Pour les enfants, nous nous sommes également intéressées à l'effet du facteur « *Age* » et à sa

[1] Les spectrogrammes ont été édités par le système UNICE Vecsys.
[2] Ces indices ont été extraits grâce à des programmes mis au point par D. Chesnet et C. Gérard du LaCo (Poitiers).
[3] L'indice est compris entre 0 et 1. Proche de 0, le contour mélodique est descendant (le pic de fréquence fondamentale est situé dans la première moitié du mot), proche de 1, le contour mélodique est montant (le pic de F0 est situé dans la deuxième moitié du mot).

possibilité d'interagir avec la « Thématisation ». Toutes les conclusions ci-dessous proviennent des ANOVAS réalisées et ne portent que sur des valeurs de F significatives.

Bien sûr, des différences interindividuelles significatives sont mises en évidence à la fois au sein du groupe des adultes et au sein du groupe des enfants. Certains individus lisent plus vite, plus fort, ou encore de manière plus aiguë que d'autres. Pour les enfants, il semble que l'*âge* soit le facteur déterminant de ces différences : la durée de lecture des syntagmes est plus importante chez les élèves de 8 ans que chez les élèves de 10 ans, leur voix est plus haute et moins forte. Par contre, aucune différence n'apparaît entre les deux groupes d'âge en ce qui concerne la gestion temporelle des pauses.

1- Niveau suprasegmental. Chez les adultes, la *thématisation* engendre un accroissement significatif des valeurs de tous les indices acoustiques considérés (durée des pauses, durée de lecture des syntagmes, intensité et fréquence fondamentale moyennes, écarts-types des valeurs d'intensité et de F0), mais pas chez les enfants. *L'interaction* entre les deux facteurs principaux (« Sujet » et « Thématisation ») n'est pas significative chez les enfants, mais l'est chez les adultes pour quatre indices sur six (durée de lecture des syntagmes, intensité et F0 moyennes, écarts-types des valeurs de F0). Le sujet 1 est le plus expressif : pour marquer l'insistance, il augmente le volume de sa voix, en élève la hauteur tonale et introduit plus de contraste intonatif que les deux autres. Le sujet 2 produit le plus fort allongement de la durée de lecture des syntagmes. Le sujet 3, quant à lui, paraît être le moins expressif puisque, bien que significatives, les modifications qu'il introduit sont les moins importantes. Ces interactions permettent donc d'isoler trois stratégies de transmission des éléments importants du texte.

2- Niveau lexical. Chez les adultes, la *thématisation* provoque une modification significative des valeurs de huit indices acoustiques sur neuf : la durée de lecture des mots-clés est allongée, l'intensité et la F0 de la voix des sujets est augmentée, les écarts-types des valeurs de F0 sont plus importants, les gammes de variation des valeurs d'intensité et des valeurs de F0 sont élargies, les pics d'intensité et de F0 ont tendance à être déplacés vers le début des mots. Chez les enfants, ce facteur n'entraîne que deux modifications significatives : la gamme de variation des valeurs de F0 est plus étendue sur les éléments thématisés et, chose curieuse par rapport aux adultes, le pic de F0 est déplacé vers la *fin* des mots. L'examen des *interactions* entre les deux facteurs principaux (« Sujet » et « Thématisation »), chez les adultes, isole encore trois stratégies différentes de transmission des informations principales du texte. Le sujet 1 augmente plus que les autres la force et la hauteur tonale de sa voix sur les éléments thématisés, il introduit des contrastes de hauteur plus marqués. Par contre, il a tendance à moins déplacer les pics d'intensité que les deux autres sujets. Le sujet 2 ralentit son élocution et introduit des contrastes d'intensité de manière plus marquée. La stratégie de transmission du sujet 3 est essentiellement basée sur le déplacement de l'accent de mot et sur le ralentissement de la vitesse de lecture. Une interaction significative entre le facteur « Age » et le facteur « Thématisation » a pu être mise en évidence : les enfants de 8 ans, tout comme les adultes, ont tendance à déplacer le pic d'énergie vers le début des mots quand ils font partie du thème principal du texte, alors que les enfants de 10 ans tendent à le déplacer vers la fin des mots.

4. DISCUSSION.

Comprendre un texte revient à construire une base macro-structurale et à élaborer des modèles de situation adaptés [Kin78]. Le phénomène de thématisation modifie la perspective de lecture et la construction du modèle de situation [Kin78]. Dans le cas de cette étude, il est légitime de penser que deux modèles de situation par texte ont été élaborés en fonction du titre présenté. Pour transmettre le produit de la thématisation à un auditeur éventuel, il faut choisir, dans le texte, les éléments les plus représentatifs du thème principal. Dans le paragraphe encadré qui développait le thème principal, la sélection des mots pertinents susceptibles de porter l'accent d'insistance a été commune à tous les adultes mais la stratégie prosodique semble être un acte propre à chaque individu. Il n'en est pas de même pour les enfants. Pour ce type de traitement particulier, les adultes semblent puiser dans un registre de configurations prosodiques stockées en mémoire à long terme [Sor89, Gér98]. Puis ils adaptent le résultat de cette recherche en fonction de leur style de parole, du contexte phrastique et textuel : un locuteur est, en effet, le premier auditeur de ses productions langagières ; des phénomènes de rétrocontrôle lui permettent de corriger les messages à tous les niveaux [Lev89]. Lire un texte à voix haute oblige les individus à une gestion fine de la prosodie sur un temps beaucoup plus long que celui de la lecture d'un mot ou d'une phrase.

Cette étude a donc mis en évidence l'existence de stratégies générales et individuelles de transmission du produit de la thématisation chez les adultes mais pas chez les enfants. Globalement, la thématisation entraîne deux grands types de modification de la parole chez les adultes : un ralentissement de la vitesse de lecture et un déplacement de l'accent de mot. L'examen des valeurs des indices de fréquence fondamentale, d'intensité et de durée de syntagmes et de mots, chez les trois adultes, a permis de montrer trois stratégies différentes. Contrairement aux adultes, qui tous jouent sur la vitesse, les données analysées chez les enfants ne montrent pas de variations temporelles dues à la thématisation, ni globalement, ni localement. De plus à 8 ans, le pic d'intensité est déplacé vers le début des mots en condition thématisée mais à 10 ans vers la fin. Une première interprétation peut être proposée,

inspirée de Thelen [The94]. Les performances des enfants dans la maîtrise d'une habileté particulière ne progressent pas de manière linéaire en fonction de l'âge. Thelen [The94], dissociant fondamentalement les performances comportementales des processus cognitifs sous-jacents, propose un modèle où l'organisme serait conçu comme un système constitué lui-même de sous-systèmes répondant chacun à des contraintes et à une trajectoire développementale propres. En fait, l'apparente « baisse » des performances à 10 ans (comparativement à l'adulte) s'expliquerait par le fait que certains sous-systèmes prendraient à un âge particulier plus d'importance en fonction des contraintes de l'environnement (qui joue le rôle d'un attracteur) ; ce qui masquerait l'efficience des autres et provoquerait la chute des performances.

BIBLIOGRAPHIE

[Clé96] Clément, J. & Gérard, C. (1996), Programmation et anticipation de l'identification des formes prosodiques. Actes des 21èmes Journées d'Etudes sur la Parole, Avignon, pp. 199-202.

[Cut80] Cutler, A. & Isard, S. D. (1980), The Production of Prosody. In B. Butterworth (Ed.) Language Production : Vol 1, Speech ans Talk. London, New York, Toronto, Sydney, San Francisco : Academic Press, pp. 245-269.

[Ead86] Eady, S. J. & Cooper, W. E. (1986) Speech Intonation and Focus Location in Matched Statements and Question, Journal of the Acoustical Society of America, 80 (2), pp. 402-415.

[Fon79] Fonagy, I. (1979), L'accent français : accent probabilitaire. In I. Fonagy & P. R. Léon (Eds.), L'accent en français contemporain. Studia Phonetica, 15. Montréal : Didier. pp. 123-227.

[Gér98] Gérard, C. & Clément, J. (1998) The Structure and Development of French Prosodic Representations, Language and Speech, 41 (2), pp. 117-142.

[Gér95] Gérard, C. & Dahan, D. (1995) Speech Durationnal Variations and Semantic Focusing in Reading, Speech Communication, 16, pp. 293-311.

[Kin78] Kintsch, W. & van Dijk, T. A. (1978) Toward a Model of Text Comprehension and Production, Psychological Review, Vol. 85, pp. 363-394.

[Lev89] Levelt, W. J. M. (1989), Speaking. From Intention to Articulation. Cambridge, Mass : The M.I.T. Press.

[Ros79] Rossi, M. (1979), Le français, une langue sans accent ? In I. Fonagy & P. R. Léon (Eds.), L'accent en français contemporain. Studia Phonetica, 15. Montréal : Didier. pp. 13-51.

[Sor89] Sorin, C. (1989), Perception de la parole continue. In M-C. Botte, G. Canévet, L. Demany & C. Sorin, Psychoacoustique et perception auditive. EM Inter/INSERM /SFA/CNET, pp. 123-139.

[The94] Thelen, E. & Smith, L. B. (1994) A Dynamic Systems Approach to the Development of Cognition and Action, Cambridge, MA : MIT Press / Bradford Books.

[Vai80] Vaissière, J. (1980), La structuration acoustique de la phrase française, Annales de l'Ecole Normale Supérieure de Pise, Vol. X, 2, pp. 529-560.

Les modalités de phrase en coréen standard : étude descriptive du contour terminal et patrons mélodiques

KIM Chongdok, YOO Hi-Yon

Université Paris 7 - UFR de Linguistique
2, place Jussieu, 75251 Paris cedex 05
Mél: chongdok@hanimail.com, hi-yon.yoo@libertysurf.fr

ABSTRACT

In Korean, the sentence ending for the familiar form [-ə] is often seen as ambiguous because there is a homonymous suffix for the four modes. However, we consider that the prosodic level eliminates any possible ambiguity. This study aims to describe the pertinent prosodic features of the terminal contour corresponding to assertive, interrogative, imperative and propositive modes of the familiar form, in production. The instrumental analysis permitted to define the parameters of F0 and duration of the sentence ending [-ə] that characterize each mode, and to establish a basic prosodic pattern for modality.

1. INTRODUCTION

En coréen, il existe des suffixes grammaticaux indiquant une fin de phrase dans lesquels sont amalgamées des informations sur :

- la modalité : assertif, interrogatif, impératif, propositif [Heo95]. Pour la modalité interrogative, nous traiterons la question totale. Quand au propositif, il correspond à la première personne du pluriel de l'impératif en français. En coréen, il s'actualise dans le registre formel par un suffixe distinct de celui de l'impératif.
- le registre : Il existe deux registres, formel et familier Selon le type de conversation et le degré de familiarité qui existe entre les interlocuteurs (par exemple, réunion de travail, discussion informelle entre amis, discours etc.), le locuteur choisira le registre le plus adéquat à la situation. Cependant, plus la situation est informelle, plus ces deux registres peuvent s'utiliser d'une manière interchangeable. Dans le registre formel, chaque modalité a un suffixe propre tandis que dans le registre familier, les suffixes qui expriment les différents types de modalité sont homonymes et se réalisent sous la forme unique [-ə].
- le degré honorifique: dans le registre familier, le degré honorifique s'exprime par un suffixe qui s'amalgame à la suite du suffixe de fin de phrase [-ə], tandis que dans le registre formel, le suffixe varie selon ce degré. On distingue deux degrés pour le registre familier (honorifique et non honorifique) et trois degrés pour le registre formel [Heo95]. L'utilisation de ces formes dépend de la relation qui existe entre les deux interlocuteurs, basée sur des critères d'âge, de statut social, de degré de familiarité etc. Certains degrés honorifiques sont interchangeables essentiellement selon la familiarité des locuteurs.

La table 1 illustre les suffixes que l'on trouve pour les formes du degré non honorifique. Dans le registre formel, on distingue deux degrés honorifiques, --honorifique étant un degré plus inférieur que le degré -honorifique.

Table 1: suffixes de fin de phrase - non honorifiques

Registre formel			Registre familier		
-hon	ass	-ne	-hon	ass	-ə
	int	-na			
	imp	-ke		int	-ə
	pro	-se			
-- hon	ass	-ninda		imp	-ə
	int	-ni			
	imp	–əra		pro	-ə
	pro	-ca			

Cette étude a comme but d'examiner le contour terminal (CT) du registre familier où l'on observe l'homonymie de la forme [-ə] pour les quatre modalités. Aucune des études récentes sur la prosodie du coréen [Jun93] [Woo93], n'aborde le problème des modalités de phrase. Quand aux grammaires traditionnelles [Kim85], elles traitent ce problème en terme d'ambiguïté. Or cette ambiguïté n'est valable qu'en l'absence de contexte syntaxique ou/et sémantique (donc en tant que phrase isolée) et en l'absence d'un contour prosodique. En effet, le niveau prosodique qui détermine un contour intonatif à la phrase efface toute trace d'ambiguïté.

En partant de l'hypothèse que les informations concernant la modalité apparaissent dans le CT qui accompagne le suffixe [-ə], nous nous proposons de faire une étude descriptive des traits prosodiques caractérisant chaque modalité et d'établir une intonation de base des modalités de phrase en coréen. Les patrons prosodiques des CT liés au suffixe comme un intonème, désambiguisent les modalités.

2. PROTOCOLE EXPÉRIMENTAL

2.1 Sujets

Nous avons enregistré six locuteurs, avec l'intention d'en éliminer un, afin d'avoir le choix des meilleurs enregistrements. Parmi les sujets, il y avait trois locuteurs féminins et trois locuteurs masculins, âgés d'environ

trente ans. Tous nos sujets sont nés et ont été élevés à Séoul et sont locuteurs natifs de ce qu'on considère le coréen standard. Ils habitent à Paris depuis une moyenne de deux ans et aucun d'entre eux n'a connu de problèmes liés à la parole.

2.2 Procédure

Corpus. Un corpus composé de huit phrases donc quatre dans le registre familier et quatre dans le registre formel a été présenté aux sujets, en écriture orthographique, sous forme de cartes et dans un ordre aléatoire. Il a été enregistré dix fois pour chaque locuteur. La première et la dernière locution ont été systématiquement enlevées et n'ont pas été prises en compte dans notre étude.

La phrase est consituée d'un complément interne <pap> ("riz"), vide de sens, d'un verbe composé du radical <mək- > ("manger") et d'un des suffixes flexionnels de fin de phrase (forme familière -honorifique et forme formelle --honorifique) présentés dans la table 1.

Table 2: corpus de l'étude

	registre formel	registre familier
assertif	pap mək-ninda	pap mək-ə
interrogatif	pap mək-ni	pap mək-ə
impératif	pap mək-əra	pap mək-ə
propositif	pap mək-ca	pap mək-ə

Son 1 : corpus registre formel par f2

Son 2 : corpus registre familier par f2

On aboutit donc à une traduction mot à mot "du riz manger" correspondant à "manger".

Enregistrement. Les enregistrements ont été menés en chambre sourde au laboratoire de phonétique de l'ILPGA de l'université Paris III, en mono, sur un minidisc SONY ES 74 minutes. La réponse de fréquence du minidisc était de 20 dB et de 40 à 11000 Hertz.

Ces enregistrements ont été ensuite numérisés avec le logiciel Sound Forge 4.5 en une résolution de16 bits et une fréquence d'échantillonnage de 22050 Hz. L'analyse et les mesures ont été réalisées avec le logiciel Winpitch 1,89 crée par Philippe Martin.

3. RÉSULTATS ET ANALYSES

3.1 Résultats

Nous considérons que la durée et la F0 sont les paramètres prosodiques pertinents permettant de caractériser les contours intonatifs de chaque modalité. Pour chaque phrase retenue (un total de huit phrases par locuteur pour chaque modalité et chaque registre) nous avons mesuré essentiellement la durée (durée totale de l'énoncé et durée du suffixe [–ə]), ainsi que la F0 de début et de fin du suffixe [–ə]. Seul pour l'impératif dont le contour est montant-descendant, une mesure a été prise à la fin de la montée: nous avons donc mesuré la durée et la F0 de la partie montante ainsi que la durée et la F0 de la partie descendante.

La durée de la voyelle, mesurée en millisecondes, a été exprimée en pourcentage par rapport à la durée totale de l'énoncé, et la F0, mesurée en Hertz, en quart de tons (QdT) sur la base de la moyenne fréquentielle du locuteur.

Les données des formes du registre formel nous ont été utiles essentiellement en terme de contrôle et de source d'explication à certains phénomènes observés dans le registre familier.

La table 3 indique les moyennes des écarts de QdT et des durées du suffixe[–ə] de chaque locuteur. La dernière colonne indique la moyenne générale des cinq locuteurs.

Table 3: Moy. des écarts de QdT et des durées (%) de [–ə] pour chaque locuteur et selon chaque modalité

			f1	f2	f3	m1	m2	M
ass	E QdT		0	1	-2	-4	0	-1
	Durée		36	29	30	30	39	33
int	E QdT		19	19	10	16	15	16
	Durée		27	35	33	30	29	33
pro	E QdT		-9	-2	-	-	-5	-5
	Durée		44	46	40	29	39	39
imp	E QdT	mon	3	3	1	3	4	3
		des	-18	-22	-11	-9	-10	-14
	Durée	mon	20	26	18	14	17	19
		des	20	19	23	22	15	20

Bien qu'il existe une certaine variabilité entre les locuteurs, l'observation de ces données permet de dégager un contour terminal spécifique pour chaque modalité, caractérisé par les deux paramètres décrits. Ainsi:

Impératif. Le contour de l'impératif se démarque des trois autres car il est le seul à avoir un contour modulé, équitablement réparti du point de vue de la durée (19% de la durée totale de la phrase pour la montée et 20% pour la descente), avec une légère montée (moyenne de 3 QdT) et une brusque descente d'une moyenne de -19 QdT. Cette chute finale de la F0 est aussi un trait caractéristique de cette modalité.

Interrogatif. Le contour de l'interrogatif se caractérise par une montée de la F0, d'environ 16 QdT. C'est la seule modalité à connaître une telle montée. Notons par ailleurs que c'est la modalité qui présente le moins de variations enter les locuteurs.

Assertif. Le contour de l'assertif est relativement plat (moyenne de -1 QdT), bien que le locuteur m1 présente un contour descendant (moyenne de -4QdT). Nous considérons que cette variation reflète un autre contour possible de l'assertif, moins important que le premier. Les données des formes du registre formel confirment ce fait, la moyenne d'écart de QdT de l'ensemble des locuteurs étant égal à –2.

Propositif. Le contour terminal du propositif est assez problématique car il présente le plus de variation inter- mais aussi intra- locuteurs. Les contours que nous avons recueillis (non reflétés dans la table 3) sont soit montants

(jusqu'à 8 QdT pour le locuteur m1) soit descendants (jusqu'à -8 QdT pour le locuteur f1).

Nous avons donc été amenés à effectuer le choix d'un contour final principal, en se basant sur la comparaison des données du registre formel. Ces dernières montrent que le contour est montant uniquement pour le locuteur m1, et systématiquement descendant pour les quatre autres locuteurs. La table 4 indique les moyennes des écarts de QdT des 4 locuteurs dont le contour est descendant.

Table 4: MOY des écarts de QdT du pro du r. formelle

		f1	f2	f3	m1	m2	M
pro	Ecart QdT	-10	-4	-2	-	-7	-6

Bien que cette décision soit arbitraire, le contour descendant reflète mieux la réalité. Nous ne rejetons pas le contour montant comme possible réalisation de cette modalité mais nous considérons le contour descendant comme le contour principal caractérisant le propositif.

3.2 Commentaires et conclusion.

La figure 1 illustre les patrons mélodiques du contour terminal des quatre modalités, déterminés par la moyenne de l'écart de QdT et de la durée de la voyelle.

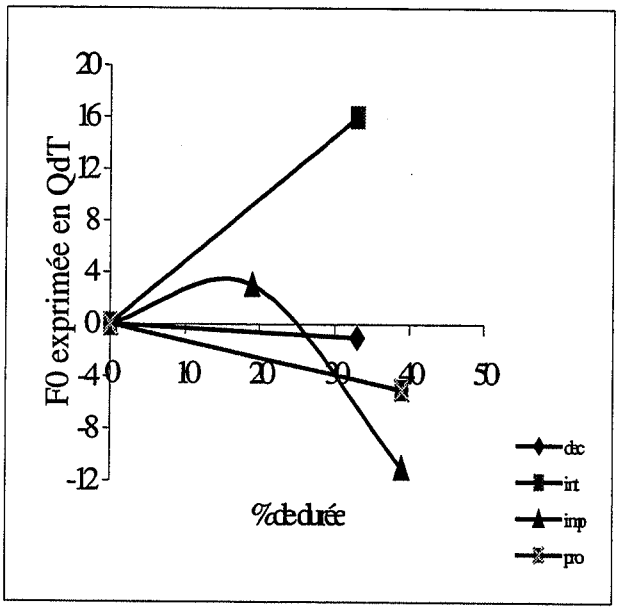


Figure 1: CT des quatre modalités avec la F0 exprimée en QdT en fonction de la durée (en %)

Dans la figure 1, il apparaît que les CT de l'interrogatif et de l'impératif se démarquent nettement, l'un pour son CT montant et l'autre pour son CT modulé montant-descendant.

Cependant, les CT de l'assertif et du propositif se ressemblent, et une possible confusion pourrait se dessiner entre ces deux CT. Il était donc nécessaire de vérifier statistiquement la différence de ces deux CT par un test-t indépendant. Le résultat obtenu pour le test-t indépendant entre les deux modalités apparaît dans la table 5:

Table 5: test-t séries non appariées pour QdT et durée

		écart moyen	t	p
ass, pro	E de QdT	4	3,9	0,0002
	Durée	-6	-4,5	<0,0001

La valeur de t étant égal à 3,9 et sa valeur de p correspondant à 0,0002 (<0.05) pour un écart de 4 QdT entre les deux types de modalités, on peut affirmer que cette différence est significative.

Par ailleurs il semble que la durée du CT de l'assertif et du propositif joue un rôle déterminant. Le résultat du test-t pour la durée, avec un t égal à -4,5 et p inférieur à 0,0001 montre que la différence est significative.

Ainsi, malgré les ressemblances entre les deux CT, les paramètres de la F0 et de la durée sont suffisamment distinctifs pour nous permettre d'affirmer ces quatre CT comme caractéristiques des modalités du coréen.

Par ailleurs, notons que les contours obtenus partagent les mêmes caractéristiques des contours des modalités en français tels qu'ils ont été décrits par Delattre [Del66] ou Martin [Mar98], et semblent confirmer ainsi l'hypothèse d'une certaine universalité de ces contours [Fon83].

4. Discussion et Conclusion

L'étude quantitative que nous avons menée sur les CT des modalités de la forme familière a permis de dégager quatre patrons mélodiques distincts caractérisant les modalités du coréen. Ceci montre donc que le terme d'ambiguïté n'est pas adéquat puisque le niveau prosodique, associé au niveau segmental, efface toute possible ambiguïté.

Nous considérons que les CT tels qu'ils ont été présentés dans la figure 1, reflètent les quatre contours mélodiques de base des modalités de phrase du coréen, d'un point de vue de la production.

Il reste à savoir si d'une part, ces CT ont également une importance significative d'un point de vue perceptif et si d'autre part, ce sont ces mêmes CT qui apparaissent dans le registre formel.

4.1 Validation perceptive

Nous avons effectué un test de perception préliminaire de CT isolés pour justifier le premier point évoqué dans le paragraphe précédent. Pour cela, nous avons choisi parmi les signaux, ceux qui correspondent au mieux aux patrons prosodiques décrits. Le suffixe [-ə] de chaque modalité a été extrait et un fichier sonore de 64 signaux (16 fois 4 modalités) représentant les CT a été crée. Six sujets, autres que les locuteurs, ont participé à ce test de perception préliminaire.

Les résultats montrent que le pourcentage de bonne réponse est à 100% pour l'interrogatif et l'impératif, sauf pour un sujet qui a confondu une fois ces deux modalités;

l'erreur peut être considérée comme non significative. Le pourcentage de bonne réponse est également de 100% pour l'assertif. Quant à la perception du CT du propositif, nous avons noté 4% d'erreur. Notons par ailleurs que quand erreur il y avait, le CT du propositif était systématiquement perçu comme un assertif. Il apparaît donc que malgré la significativité de la différence entre ces deux modalités, une confusion est parfois possible. Toutefois, on pourrait espérer qu'un test de perception fait à partir des stimuli de synthèse, où tous les paramètres seraient contrôlés, aurait donné de parfaits résultats.

4.2 Comparaison avec la forme formelle

Nous avons vu qu'il existe en coréen deux formes de suffixe de fin de phrase, s'utilisant selon la situation d'énonciation. Or, tandis que le CT de chaque modalité se réalise d'une manière assez homogène dans le registre familier, il se trouve que les formes du registre formel présentent une grande variation, non seulement inter-locuteurs, mais aussi intra-locuteur. Par conséquent, pour le registre formel, il nous a été difficile d'associer, d'une manière univoque, un patron mélodique du CT de fin de phrase à une modalité. L'observation des données de la forme formelle nous permet de faire les remarques suivantes d'ordre qualitatif :

- Dans la section 3, nous avons montré les variations du propositif du registre formel; c'est d'ailleurs sur la comparaison des deux registres que se base le CT choisi pour cette modalité.

- Le CT de l'interrogatif présente, comme pour les données du registre familier, une absence de variation: nous observons une montée, dont la pente est, toutefois, plus ou moins brusque.

- Le CT de l'assertif et celui de l'impératif présentent le plus de variation. On observe des contours montants, descendants, plats et modulés, ces variations pouvant apparaître chez un même locuteur. Les variations sont telles qu'il est difficile d'établir le choix d'un contour proéminent.

Une possible explication de ces nombreuses variations des CT dans le registre formel, peut venir de la relation qui existe entre le niveau prosodique et le niveau syntaxique. Il semblerait que la présence d'un suffixe spécifique à la modalité permet d'atténuer la conservation du CT de fin de phrase et de présenter ainsi des variations. En effet, le suffixe se suffisant en lui-même pour indiquer le type de modalité, le niveau prosodique peut prendre une autre fonction, attribuée par les informations paralinguistiques (attitudes, émotions etc.). Nous considérons que les CT qui s'écartent des intonations de base des modalités, ne sont pas dépourvus de ces données paralinguistiques.

Quant aux formes du registre familier, l'homonymie des suffixes contraint le niveau prosodique à préserver le CT correspondant à la modalité en question. Dans ce cas, les deux niveaux doivent rester étroitement liés, en ne laissant qu'une petite marge pour une possible variation.

4.3 Conclusion

Nous avons mené une étude quantitative des paramètres de la F0 et de la durée afin de décrire le contour terminal des quatre modalités de phrase du coréen dans le registre familier. Cette étude a permis de dégager un contour terminal typique à chaque modalité (cf. figure 1) et d'écarter la possibilité d'une ambiguïté entre ces phrases. Nous avons ainsi déterminé le rôle proéminent du niveau prosodique, nécessaire à l'interprétation de ces dernières. Par ailleurs, ces CT déterminés par la production, semblent être tout aussi pertinents au niveau perceptif mais une étude perceptive plus détaillée devrait confirmer cette hypothèse. D'autre part, reste le problème du lien exact des CT du registre familier et du registre formel, afin de donner une explication plus rigide aux nombreuses variations rencontrées dans le registre formel. Enfin, il serait intéressant de compléter cette étude en examinant les contours terminaux des suffixes du registre familier et du registre formel qui varient selon le degré honorifique afin de valider les contours terminaux que nous avons dégagés pour chaque modalité.

BIBLIOGRAPHIE

[Del66] Delattre P.(1966), "Les dix intonations de base du français", French Review, Vol.40(1), pp.1-14.

[Fon83] Fonagy I. (1983), La Vive Voix, Payot.

[Heo95] Heo W. (1995), Morphologie du coréen du 20ème siècle, Sam Munhwasa.

[Jun93] Jun S-A. (1993), The Phonetic and Phonology of Korean Prosody, PhD dissertation, Ohio, Ohio University.

[Kim85] Kim S-D, Kim C-K, Lee K-M. (1985), Phonologie du coréen, Presses de l'Université de Hankook Pangsong Thongsin

[Mar98] Martin P. (1998), "Systèmes prosodiques et énonciation", Gragoatá Revista do Instituto de Letras, Niteroï, RJ, Brésil pp.21-40.

[Woo93] Woo-Baluc L-S. (1993), "Accent primaire, accent secondaire et patrons rythmiques en coréen standard lu", TiPS, 23, 95-154.

Variations tonales et structure prosodique de la focalisation en somali.

Le Gac David

Université Paris 7- UFR de Linguistique

2, place Jussieu, 75251 Paris cedex 05

Mél: legac@ccr.jussieu.fr

ABSTRACT

In this paper, I investigate the phonetic and phonological markers of two categories of nouns (N1, N2) in [±focus] and [±final] positions in Somali, a tonal-accent language of the cushitic familly. I found that 1) in [-Foc] position, N1 and N2 have a final tone ; 2) in [+Foc] position, the tone of N1 has the same place as in 1), the one of N2 is also final in [-Fin] position, but is penultimate in [+Fin] position, and 3) the focus triggers a specific intonational structure. Adopting the frameworks of [P&B88] and [Mar81], I capture all these variations, and I propose a definition of tonal-accent language.

1. INTRODUCTION

Dans cet article, j'étudie un phénomène très courant à travers les langues : la focalisation. Phonétiquement, le focus est marqué par une proéminence spécifique et une restructuration prosodique de l'énoncé.

Ainsi, en français, il est marqué par un contour terminal (CC) et éventuellement par un accent sur la première syllabe [Ros99]. En suédois, le focus assigne un accent de phrase [Bru77,Tou87]. En japonais, le mot focalisé a un accent tonal proéminent dans la phrase [P&B88]. En outre, dans toutes ces langues, les éléments en préfocus contrastent avec ceux en postfocus. Typiquement, le focus "écrase" ou efface les accents post-focaux.

En ce qui concerne le somali, le focus est marqué syntaxiquement par une série de morphèmes : *ayaa* ou *baa*, qui focalisent un item qui les précède immédiatement, et *waxaa* qui focalise un élément en fin de phrase [Sae98]. On a donc pensé qu'en somali, la focalisation n'était pas pris en charge par l'intonation car celle-ci aurait été redondante [Sae98, Hym81]. En outre, le somali est une langue à accent tonal (dorénavant AT), renforçant cet *a priori* [id.]. Or quatre raisons vont à l'encontre de cette idée : 1) dans des langues à AT, comme le suédois ou le japonais, le focus est pris en charge par l'intonation ; 2) l'existence de morphèmes spécifiques n'empêchent pas le focus d'être marqué tonalement [Ros99] ; 3) il existe en somali, un phénomène de "Downdrift", or celui-ci est un paramètre intonatif reconnu [Bru77], et finalement, 4) un patron tonal spécial semble bien apparaître sur une classe de noms focalisés.

En effet, d'après Andrzejewski [And64], il existe deux catégories de noms (N1 et N2) dont le comportement tonal diffère : en position [-Foc], N1 et N2 ont un AT final ; en position [+Foc], N1 a également un AT final, mais N2 présente une asymétrie : l'AT est final en position [-Fin] de phrase, mais pénultième en [+Fin]. Il faut noter que seul [And64] mentionne ce fait.

En somme, le but de ce travail est 1) de vérifier instrumentalement les données décrites par [And64], 2) de montrer que les variations tonales entre N1 et N2 procèdent d'un même mécanisme phonologique où entre en jeu un ton B qui s'associe à la tête prosodique de l'énoncé, et 3) que la focalisation en somali est bien prise en charge par l'intonation, et qu'en outre, celle-ci a une structure semblable à d'autres langues. Autrement dit, je montrerai qu'il existe également un contraste entre les éléments pré- et post-focaux, et que le ton B postulé en 2) semble apparaître dans d'autres langues.

2. EXPÉRIENCES ET RÉSULTATS

2.1 Méthode

Deux types de phrase ont été utilisés (table1) :

1) dans la première série, le N[+Foc] précède le focalisateur *ayaa*, et n'est donc pas en finale d'énoncé, tandis que le N[-Foc] est en fin de phrase,

2) dans la deuxième série, à l'inverse, N[+Foc] est en fin de phrase. Le N[-Foc] se situe juste avant *waxaa*.

Un nom de chaque catégorie a été placé dans chacun de ces contextes. Les autres éléments sont les mêmes dans chaque type de phrase : un adverbe en début d'énoncé, et un GV après le focalisateur *ayaa* ou *waxaa*.

Table 1 : corpus. S11 et S22 : "*C'est une fille qui a parlé à Cige*". S12 et S21 : "*C'est Cige qui a parlé à une fille*".

Séries	Adv	N[-Fin]	Focal	GV	N[+Fin]
S1		N[+Foc]	ayaa		N[-Foc]
S11	Shaley	gabadh	ayaa	la hadashay	Cige
	Hier	*une fille (N1)*		*elle a parlé*	*(à) Cige (N2)*
S12	Shaley	Cige	ayaa	la hadlay	gabadh
	Hier	*Cige (N2)*		*il a parlé*	*(à) une fille (N1)*
S2		N[-Foc]	waxaa		N[+Foc]
S21	Shaley	gabadh	waxaa	la hadlay	Cige
	Hier	*(à) une fille (N1)*		*il a parlé*	*Cige (N2)*
S22	Shaley	Cige	waxaa	la hadashay	gabadh
	Hier	*(à) Cige (N2)*		*elle a parlé*	*une fille (N1)*

Chaque phrase a été produite 5 fois, et répondait à une question du type : "Hier, qui a parlé avec X, c'est Y ? Non, → réponse", afin de provoquer la focalisation.

Questions et réponses ont été écrites sur des fiches

individuelles, mélangées et présentées aléatoirement au locuteur qui les lisait à un débit normal. L'enregistrement et l'analyse ont été effectués à l'aide du logiciel WinPitch1.87. Malheureusement, je n'ai pu bénéficier que d'un seul locuteur jusqu'à présent.

2.2 Résultats

Les figures 1 et 2 présentent les courbes de f0 des énoncés S1 et S2. Une mesure a été prise sur chaque syllabe, et sur les segments présentant un minimum local de f0 (ex.: le [w] de *waxaa*).

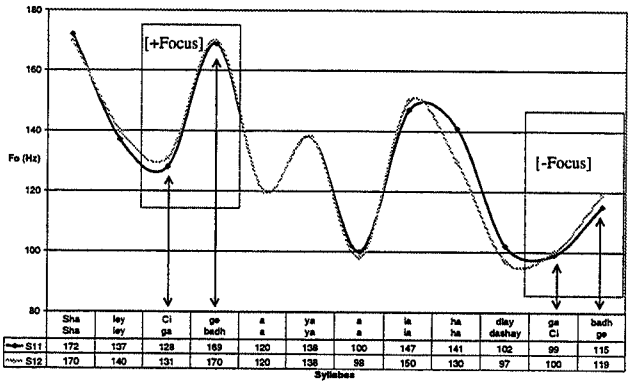


Figure1 : moyennes de f0 des énoncés S1

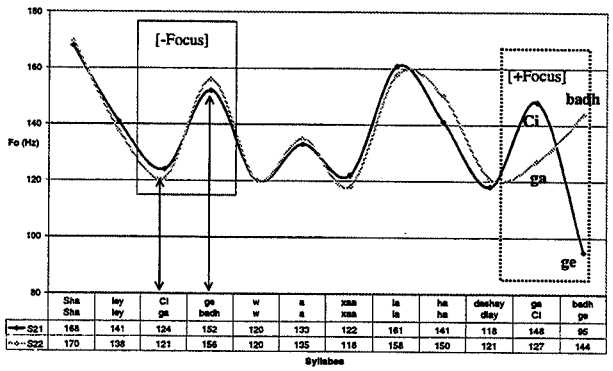


Figure2 : moyennes de f0 des énoncés S2

On voit qu'un AT se réalise en une montée de f0 dont le maximum est atteint sur la syllabe accentuée et le minimum, sur la syllabe précédente.

Mes résultats confirment les observations de [And64], à savoir que N2[+Foc] en final de phrase a un AT pénultième (figure 2), alors que tous les autres AT sont finaux.

Les figures 1 et 2 montrent clairement que le somali a une structure intonative déclenchée par la focalisation. En outre, elle est différente selon la place du focus. Si l'on observe la hauteur des AT, on remarque les faits suivants :

a) la valeur du H de l'adverbe est stable (170Hz), et constitue le point le plus haut des énoncés S2,

b) quand le N[-Fin] est non-focalisé (figure2), le Downdrift (dorénavant DD) s'applique jusqu'à l'AT du focalisateur *waxaa*. Par contre, le DD ne s'applique pas sur le N focalisé (figure1), lequel demeure à la même hauteur que l'adverbe (170Hz).

c) Dans les deux séries, le GV est plus haut que le focalisateur, et, dans la S2, plus haut que le N[-Fin][-Foc]. Il reste toutefois moins important que l'AT de l'adverbe (S1/S2) et du N[-Fin][+Foc] (S1). Il y a donc un contraste entre la hauteur du GV selon la place du focus : quand le N[+Foc] précède le GV, celui-ci est à 149Hz en moyenne, mais lorsque le N[+Foc] est final, le GV culmine à 160Hz.

d) Enfin, la hauteur du N[+Fin] dépend aussi de la focalisation : le N final est plus haut quand il est focalisé (146Hz *vs* 117Hz en moyenne). Par ailleurs, il semble abaissé par rapport à ce que prévoit le DD (en S1 : 117Hz au lieu de 130-135Hz)

Quant aux jonctures, elles se réalisent comme un ton B. Elles sont relativement stables, sauf en S1 où le ton B qui suit le focalisateur *ayaa* atteint le minimum de l'énoncé.

En somme, les items qui suivent le focus sont moins hauts que lorsque ces mêmes items précèdent le focus. Cette configuration se retrouve dans d'autres langues à AT : en suédois [Bru77, Tou87] ou en japonais [P&B88], le focus abaisse les items qui le suivent, sans toutefois effacer leurs accents comme en français. Le somali a donc bien une structure intonative de langue à accent tonal.

Par ailleurs, on retrouve en somali un fait qui semble apparaître également dans les langues mentionnées ci-dessus : un ton relativement bas par rapport aux autres jonctures, se plaçant après le focus, soit à la fin du focalisateur *ayaa*, soit à la fin de la phrase, précisément là où l'AT de N2 varie.

3. ANALYSE ET MODÈLE

3.1 Représentation de l'accent tonal

Analyses ultérieures. [Hym81] a proposé de représenter l'accent tonal comme l'association d'un accent sous-jacent, symbolisé par *, à un ton H. Cela signifie que la présence, au niveau phonétique, d'un ton H implique la présence d'un accent sous-jacent et *vice et versa*.

Or, l'observation des tons H montre que 1) le domaine de réalisation d'un ton H accentuel est la more vocalique, 2) on ne trouve pas de H au-delà de la more vocalique pénultième, et 3) il n'y a qu'un H par mot. Hyman a donc proposé les 3 règles suivantes pour l'assignation des * au mot : 1) les * sont assignés à la more vocalique (dorénavant μ) ; 2) un * ne peut être associé au-delà de la μ pénultième ; 3) il ne peut y avoir qu'un seul * par mot

Cependant, si les deux premières règles sont acceptables, la troisième pose un problème. En effet, postuler un seul * par mot amène à poser la structure accentuelle suivante pour N1 et N2 : [...°*]. Or, en position [+Fin][+Foc], il faut prévoir pour N2 une règle de déplacement accentuel (°*→*°), alors que pour N1, cette règle ne jouerait pas. La règle 3) amène donc à poser une règle *ad-hoc*. Bref, la règle 3 empêche toute explication cohérente des variations tonales en somali. Je propose de la rejetter.

Le rejet de la règle 3) implique qu'en somali, un mot peut

avoir deux accents sous-jacents, à la fois sur la μ finale et sur la μ pénultième, mais pas plus en respect des deux premières règles.

Or, c'est ce que suggère en fait le comportement tonal de N2. Le fait que celui-ci puisse avoir un AT sur la μ finale ou pénultième indiquerait donc qu'il a deux *. Par contre, le caractère fixe de l'AT de N1 indiquerait que ces noms n'ont qu'un * sur la μ finale. On va voir que cette hypothèse permet d'expliquer la tonologie du somali.

Toutefois, selon la conception traditionnelle de l'AT, où à chaque * correspond un ton H, N2 devrait avoir deux tons H. Il faut donc expliquer pourquoi N2 n'en a qu'un seul.

Le cadre théorique de [P&B88] Le type de représentation proposé par [P&B88] fournit une explication à ce problème.

Dans ce modèle, les tons sont affectés à un constituant prosodique, soit à sa droite, soit à sa gauche. Les tons s'associent ensuite aux Unités Porteuses de Tons (UPT) qui sont en marge, ou qui sont têtes des constituants.

En somali, je propose qu'il y ait un constituant "mot" (noté ω), auquel est affecté à droite un ton H. Celui-ci s'associe de droite à gauche, à une μ qui porte un *. La figure 3 montre ce mécanisme. Ainsi chaque mot n'a qu'un seul ton H, même si N2 a deux *.

Figure 3. Représentation de l'AT

Ce type de représentation permet par ailleurs de cerner d'un manière plus formelle ce qu'est une langue à AT : il s'agit d'une langue où un ton accentuel appartient au domaine du mot. Une langue à accent serait alors une langue où aucun ton ne serait affecté au constituant mot.

Cependant, on a vu que le H de N[+Fin][+Foc] se déplace sur sa μ pénultième. Dans la partie suivante, je propose une explication à ce phénomène.

3.2 Déplacement du H de N2

Un premier élément de réponse vient de la représentation donnée dans la figure 3. Je rappelle que les UPT auxquelles doit s'associer H sont les μ affectées d'un *. S'il est possible pour H de N2 de se déplacer, c'est parce que justement la μ pénultième de N2 porte un *, alors que la μ pénultième de N1 n'en porte pas. Autrement dit un ton H peut se déplacer si une autre μ* lui est accessible.

Il reste toutefois à expliquer la cause du déplacement du H lorsque N2 est focalisé et en fin de phrase, alors que tous les autres H restent finaux.

Dans les figures 1 et 2, le focus est suivi d'un ton B atteignant le minimum du locuteur. Ce ton B prenait place sur la dernière μ du focalisateur *ayaa*, mais on le retrouvait également en fin de phrase, précisément là où N2 a un accent pénultième. Je propose donc l'hypothèse suivante : le déplacement du H de N2 est dû à ce ton B, lequel marque la focalisation. Ce ton vient de la droite et s'associe à une UPT en marge droite de constituant.

Ainsi, le B de focus désassocie le H de N2 car celui-ci peut se déplacer sur la μ pénultième qui porte un *. Mais il reste flottant en N1 car le H ne peut se réassocier (figure 4).

Figure 4

Un problème demeure : les N2 devant *ayaa* gardent un AT final alors qu'on s'attend à ce que le B de focus viennent le déplacer sur la μ pénultième. En fait, la figure1 montre clairement que le ton B de focus s'associe à la dernière more de *ayaa* et non pas à celle de N2.

Pour expliquer pourquoi B ne "remonte" pas jusqu'au mot focalisé, je propose l'hypothèse suivante : le nom et *ayaa* appartiennent au même groupe prosodique auquel est associé B de focus. Il existe au moins trois raisons pour regrouper ces deux items dans un même constituant : 1) une raison syntaxique : *ayaa* focalise l'élément qui le précède immédiatement, il est donc naturel de les voir réunis au sein d'un même constituant ; 2) une raison phonologique : le focalisateur *baa*, strictement équivalent à *ayaa*, peut se suffixer au N qui précède en même temps que le *b-* tombe (*gabádh+báa* > *gabádháa*) ; et 3) une raison prosodique : le H du focalisateur est abaissé par le nom qui précède.

Dans cette section, j'ai donc proposé une explication de la variation tonale de N2. Dans la partie suivante, je montre à quel type de constituant est affecté le B de focus, en même temps que je présente un modèle qui rend compte de la structure intonative du somali.

3.3. Structure prosodique

Le fait que N et le focalisateur soient dans un même groupe indique qu'il existe un constituant supérieur au mot. Je l'appellerai "focus prosodique" (noté φ), et propose que le ton B de focus lui soit associé à droite.

Par ailleurs, la hauteur maximum du N[-Fin] indique que φ est la tête prosodique de l'énoncé. De plus, cela paraît naturel pour un constituant qui porte la focalisation.

Ainsi le ton B de focus est affecté au constituant prosodique tête de l'énoncé, d'où le minimum fréquentiel qu'il atteint. Un fait indépendant vient confirmer cette hypothèse. En isolation, N2 a également un AT pénultième [And64,Hym81]. Or, un mot en isolation ne peut être que sa propre tête. Il semble donc que la fonction de B soit de marquer un constituant tête. Il est remarquable que l'on ait un fait similaire en français : un nom focalisé ou en isolation est également marqué par un contour terminal, *i.e.* un ton B [Ros99]. Bref, on est peut-être en présence d'une caractéristique intonative largement répandue.

Poser l'existence d'une tête prosodique implique une structure hiérarchique dans laquelle prennent place les

autres constituants. Pour dégager une telle structure, je me suis fondé sur les hauteurs relatives des tons H (Figure 5).

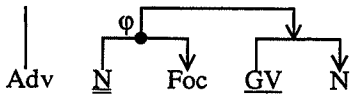


Figure5 : structure prosodique des phrases S1.

A l'instar de la grammaire générative, on peut poser des relations de gouvernement entre les constituants têtes (soulignés) et leur "complément". Cette relation se matérialise par l'abaissement phonétique du ton H de l'élément gouverné, c'est-à-dire le downdrift, elle est donc orientée de gauche à droite.

On remarque que le N[+Foc] gouverne le GV. Autrement dit, cette représentation montre clairement que le DD n'est pas une règle phonétique s'appliquant entre deux éléments adjacents en surface, comme le proposent [P&B88], mais bien un indicateur de relations dans une structure phonologique.

L'adverbe est à la même hauteur que le N[+Foc] qui suit, il n'est donc pas gouverné par lui, mais il ne le gouverne pas non plus, ce qui paraît normal si l'on considère que N est la tête de l'énoncé. Bref, le focus ne gouverne pas l'élément à sa gauche, lequel se retrouve finalement hors structure (dorénavant HS).

On retrouve le même type de structure proposée par [Mar81] pour le français, à ceci près que dans cette langue le "gouvernement" s'applique de droite à gauche, et qu'il se traduit phonétiquement par l'inversion de pente. Un constituant post-focus sera donc HS, et se réalisera sans contours, dans le registre bas. En somme, le somali semble être une "image miroir" du français. Mais fondamentalement, un mécanisme commun sous-tend les deux langues.

Un autre fait vient conforter cette idée : le fait d'être HS implique la non-spécification intonative, l'item HS va alors "copier" la spécification tonale du premier élément gouverneur. Ainsi en français, l'item post-focus est bas car il copie le ton B terminal du focus [Ros99]. En somali, on peut postuler le même principe : l'Adv est à la même hauteur que le N[+Foc] car il copie la spécification tonale du focus qui est [hauteur maximale de l'énoncé].

On s'attend alors à ce que les éléments pré-focus soient à la même hauteur que le focus dans les phrases S2. Or, la figure 2 montre que ce n'est pas le cas, et indique plutôt la structure prosodique suivante (figure6) :

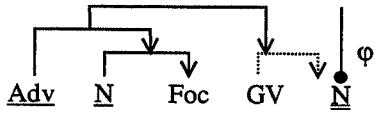


Figure 6 : structure prosodique des phrases S2

Deux problèmes se posent : 1) le N[+Fin][+Foc] n'est pas à la hauteur maximale de l'énoncé, et semble même être gouverné par le GV ; 2) le DD s'applique entre les éléments pré-focus. Cependant :

1) En suédois [Tou87], le dernier élément d'une phrase est abaissé, certainement pour des raisons physiologiques. En somali, on retrouve cet abaissement dans les phrases S1 (N[+Fin] à 117Hz au lieu de 130-135Hz). Autrement dit, en S2, le focus final est **phonologiquement** à 170Hz, et demeure la tête de l'énoncé. Le gouvernement du N[+Foc] par le GV ne serait donc qu'une "illusion d'optique".

2) La figure 6 montre que l'Adv gouverne les éléments pré-focus. Il est donc tête d'une structure regroupant les items pré-focus. Cela implique qu'au niveau le plus haut de la structure de l'énoncé, Foc est adjacent à N[+Foc]. C'est donc lui qui copie la spécification prosodique du focus. C'est ce que montre en fait la figure 6 : Adv est également à 170Hz.

En somme, une structure de ce type permet de rendre compte de l'intonation du somali, mais également de celle du français. La différence entre les deux langues réside dans l'orientation du gouvernement, avec la conséquence suivante : l'une est l'image miroir de l'autre. On pourrait pousser la généralisation plus loin, en proposant l'hypothèse suivante : la structure prosodique d'une langue à AT (somali, suédois, japonais...) est l'image miroir de la structure prosodique d'une langue à "intonation" (français, anglais).

4. Conclusion

Dans cet article, j'ai rendu compte de la phénoménologie tonale du somali, en proposant plusieurs hypothèses : 1) N2 a deux *, auxquels s'associe un seul ton H affecté au mot, 2) le déplacement de l'AT de N2[+Fin][+Foc] est dû à un ton B marquant le constituant prosodique tête de l'énoncé, et qui semble apparaître avec la même fonction en français, 4) enfin, j'ai montré que l'intonation du somali était prise en charge par une structure prosodique, qui serait "l'image miroir" de celle du français.

Bibliographie

[Ros99] Rossi M. (1999), L'intonation, le système du français, Ophrys

[Bru77] Bruce G. (1977), Swedish Word Accents in Sentence perspective, Lund, Gleerup

[Tou87] Touati P. (1987), Structure prosodique du suédois et du français, Lund University Press

[P&B88] Pierrehumbert J. & Beckman M. (1988), Japanese Tone Structure, MIT Press

[Sae98] Saeed J.I. (1998), Somali, CLCS, Dublin

[Hym81] Hyman L. (1981), "Tonal Accent in Somali", Studies in African Linguistics,V12,pp.169-201

[And64] Andrzejewski B.W. (1964), The Declensions of Somali Nouns, University of London

[Mar81] Martin Ph. (1981), "Pour une théorie de l'intonation", in *Rossi et al.*, Klincksiek, pp. 234-271

Effets articulatoires de l'emphase contrastive sur la Phrase Accentuelle en français

Hélène Lœvenbruck

Institut de la Communication Parlée (ICP) – INPG/ Univ. Stendhal/ UMR CNRS 5009 –
46 avenue F. Viallet – 38031 Grenoble – France – Mél : loeven@icp.inpg.fr

ABSTRACT

Articulatory effects of contrastive emphasis on the *Accentual Phrase* (AP) in French are considered here. In [Jun95]'s model, the AP features an initial high tone Hi, ('accent secondaire'), a final high tone H* ('primaire'), and 2 low L tones preceding them. Sentences with 4-syllable APs were recorded for 2 French speakers, using EMA. The position of the AP in the sentence varied, several speaking conditions were elicited. Vertical displacement, peak velocity and movement duration were analyzed for a pellet on the tongue middle. A preliminary study [Lœv99] on 1 subject suggested that LHi could be related to hyper-articulation of the first syllable, and LH* to even stronger hyper-articulation of the last syllable. These results are confirmed for a second speaker, and the effects of emphasis depending on the speaker are studied.

1. INTRODUCTION

Des travaux récents [Bec96] montrent que la prosodie est une structure linguistique hiérarchique à part entière et qu'il est impératif de mieux s'entendre sur ses caractéristiques phonologiques et phonétiques. Les études articulatoires de la prosodie du français (e.g. [Vat93]) fournissent des conclusions variées. Les irrégularités observées pourraient être dues au fait que la structure prosodique est rarement prise en compte et que des phénomènes différents sont regroupés. Cette étude examine les effets articulatoires de l'emphase sur une entité prosodique, la Phrase Accentuelle, en s'appuyant sur un modèle linguistique de la prosodie du français.

Jun & Fougeron [Jun95, Fou98] ont proposé un modèle de l'intonation du français dans lequel le niveau le plus bas est la Phrase Accentuelle (AP) et le plus élevé est la Phrase Intonationnelle (IP). L'IP est marquée soit par une montée de continuation finale majeure (H%), soit par une chute finale majeure (L%) ainsi que par un allongement final éventuellement suivi d'une pause. L'AP, située au-dessus de l'unité tonale de Hirst & Di Cristo [Hir98], correspond au « mot prosodique » de Vaissière [Vai92], à l'« intonation group » de Mertens [Mer93], l'« intonème mineur » de Rossi [Ros85]. Sa représentation tonale sous-jacente est /LHiLH*/, avec un ton haut initial Hi (aussi nommé sommet de « l'accent secondaire »), un ton haut final H* réalisé sur la dernière syllabe pleine (sommet de « l'accent primaire ») et deux tons bas L réalisés sur les syllabes précédant les syllabes portant un ton H. Notre étude porte sur les caractéristiques articulatoires de l'AP.

Dans une étude préliminaire [Lœv99], nous avons analysé et comparé les caractéristiques articulatoires et acoustiques des syllabes de plusieurs mot-cibles. Ces mots-cibles, composés de 4 syllabes, étaient susceptibles de constituer une AP. D'après la définition ci-dessus, notre prédiction était donc d'observer un accent haut H* sur la dernière syllabe et un accent haut Hi sur la 2ème syllabe. Le corpus avait été construit pour faciliter l'étude acoustique et articulatoire. Ainsi, pour permettre le suivi de F0, seules des sonores ont été utilisées et pour l'analyse des mouvements de la langue, les voyelles /i/ et /a/ ont été préférées. Les mots-cibles suivants ont donc été choisis : l'anonymat, l'anomala, l'éliminé, l'illuminé, l'aluminium, l'ignominie, l'inaliénée, l'énamouré, l'inanimé. Ces mots-cibles, prononcés par une locutrice française, ont été insérés dans des phrases intonationnelles, en 3 positions (initiale, centrale et finale), comme par exemple :

'L'illuminée a allumé néanmoins le monument.' (Initial).
'Il a humilié l'illuminée en l'éloignant.' (Central).
'L'aumônier a néanmoins éloigné l'illuminée.' (Final).

Du point de vue acoustique, l'étude préliminaire a montré que le contour /LHiLH*/ est généralement bien observé lorsque l'AP est en positions initiale et centrale dans l'IP et, conformément au modèle linguistique, remplacé par [LHiLL%] en position finale. L'alignement de LH* est conforme aux prédictions, H* portant sur la dernière syllabe et L sur la syllabe précédente. Cependant, comme il a souvent été noté dans la littérature, l'alignement de Hi est variable, depuis la première à la deuxième syllabe, la première syllabe étant parfois porteuse d'une montée de F0 atteignant son sommet sur la syllabe suivante. Du point de vue articulatoire, l'étude semblait indiquer que, pour cette locutrice, l'accent secondaire (ou LHi) pourrait être relié à une légère hyper-articulation de la 1ère ou de la 2ème syllabe de l'AP. L'accent primaire (ou LH*) à une hyper-articulation de la dernière syllabe.

L'objectif de cet article est de valider ces premiers résultats pour un deuxième locuteur et d'étudier les effets de l'emphase sur les caractéristiques articulatoires de l'AP chez deux locuteurs.

2. MÉTHODE EXPÉRIMENTALE

2.1 Corpus

Le corpus consiste en des phrases intonationnelles, contenant le mot-cible (l'AP) « l'anomala », en 3 positions (initiale, centrale, finale) dans l'IP. L'analyse porte sur la première et la dernière syllabe de l'AP : σ1 et σ4.

Les phrases ont été enregistrées dans 2 « modes ». En **mode naturel :** les phrases sont simplement lues le plus naturellement possible. En **mode contrastif :** avant chaque enregistrement, les locuteurs entendent la phrase à prononcer, dans laquelle le mot-cible a été remplacé par un autre. Ils ont pour tâche de corriger la phrase, en plaçant un accent d'emphase contrastive sur le mot-cible.

2.2 Sujets et enregistrements

Les signaux articulatoires et acoustiques ont été enregistrés simultanément, en chambre sourde, pour 2 locuteurs français (une femme AV, un homme SL), en utilisant l'articulographe de l'ICP (EMA Cartens AG100). Cinq bobines ont été collées dans le plan sagittal : en 3 points de la langue (apex, milieu, dos) et sur les incisives inférieure et supérieure (point de référence). Les données EMA ont été échantillonnées à 500Hz, le signal acoustique à 16000Hz.

Dans chacun des 2 modes, 3 conditions d'élocution ont été enregistrées. Pour AV, dans le mode naturel, 3 niveaux de clarté ont été mis en œuvre : *normal* – l'instruction était de parler avec un débit et une articulation confortables–, *clair* –l'instruction était d'être claire pour l'auditeur–, *très clair* –être encore plus claire. La consigne de clarté n'étant pas aisée à mettre en place, nous avons préféré une consigne de débit pour le deuxième locuteur. Trois niveaux de débit ont donc été enregistrés : *rapide*, *normal*, *lent*. En mode contrastif, 3 niveaux de débit ont été enregistrés, pour les 2 locuteurs.

Outre le corpus, un enregistrement de calibrage a été nécessaire pour effectuer le tracé du contour du palais de chaque locuteur. Puis, afin de corriger la rotation des données dans le plan sagittal, l'angle entre l'arrière du palais et l'axe horizontal a été mesuré pour chaque sujet. Les tracés articulatoires ont alors subi une rotation appropriée, de sorte que le nouvel axe des abscisses corresponde à la dimension horizontale du locuteur et l'axe des ordonnées à sa dimension verticale. Les données articulatoires ont ensuite été filtrées par un filtre passe-bas et normalisées par la bobine de référence pour compenser les mouvements de la tête. Les vitesses et accélérations ont été obtenues à partir des positions enregistrées en utilisant une méthode aux différences finies.

3. RÉSULTATS ET DISCUSSION

3.1 Contours de la fréquence fondamentale

Nous avons d'abord vérifié que les contours de F0 suivaient le motif décrit par [Jun95]. Les signaux acoustiques étant de mauvaise qualité, il n'a pas toujours été possible de bien suivre les contours et une écoute attentive a été nécessaire pour l'analyse mélodique. Pour AV, le contour LHiLH* est bien observé (remplacé par LHiLL% en fin d'IP), Hi étant généralement aligné avec σ1 et H* toujours avec σ4 (cf. figure 1). Si l'alignement de Hi avec σ1 ou σ2 est quelques rares fois difficile à spécifier en mode naturel, il porte toujours sur σ1 sous emphase. Pour SL, l'alignement de Hi est plus variable, même en mode contrastif. Cependant, lorsque le sommet de Hi est aligné avec σ2, σ1 porte une montée de F0. Les débats abondent sur la position de l'accent secondaire. Nous émettons l'hypothèse que LHi est ancré à gauche mais de façon « libre », les variations seraient donc un artefact de l'alignement tonal d'un accent non-associé.

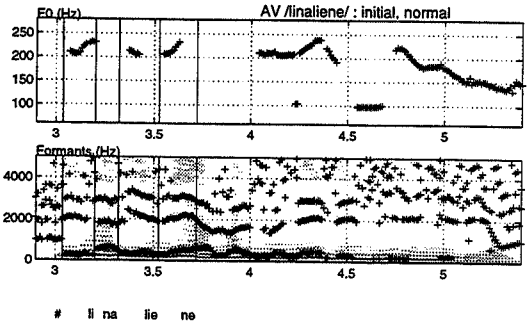


Figure 1 : Suivi de F0 et formants pour l'IP "L'inaliéné a malmené l'ennemi humiliant.", condition normale, mode naturel, locuteur AV. LHi aligné sur σ1, H* sur σ4.

3.2 Mouvements de la langue

L'analyse globale des 8 trajectoires pour les 4 bobines (positions horizontales et verticales de l'incisive inférieure et de l'apex, du milieu et du dos de la langue) nous a conduite à sélectionner la position verticale de la bobine située sur le milieu de la langue (Ymili), celle-ci présentant le plus de variations et étant la plus lisible pour l'articulation du /la/.

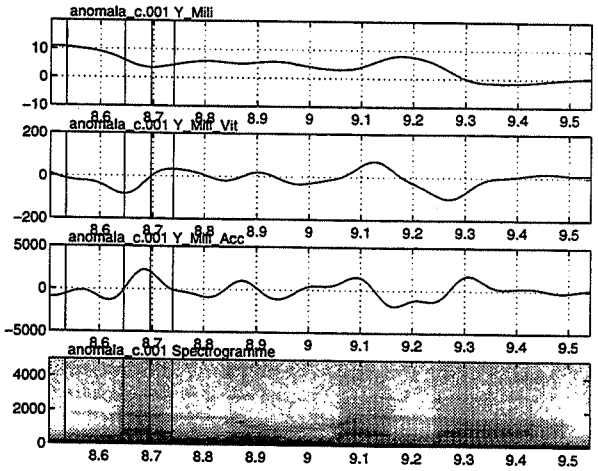


Figure 2 : Etiquettes articulatoires pour σ1 de [lanomala] dans 'Elle annihilait l'anomala en l'éloignant'. Premier panneau : position verticale de la bobine située au milieu de la langue (en mm) ; second panneau : vitesse (mm/s) ; troisième panneau : accélération (mm/s^2).

Pour chaque syllabe, nous avons mesuré 3 paramètres :

- le déplacement d'une position de référence pour /l/ à la position où la voyelle /a/ était pleinement atteinte,
- le pic de vitesse pour le mouvement de /l/ à /a/,
- et la durée totale de la syllabe.

Comme le montre la figure 2 pour [la-no-ma-la], le début et la fin de la syllabe (qui fournissent la durée) ont été repérés sur le sonagramme ; le déplacement maximum de /l/ à /a/ a été repéré par le passage par zéro de la vitesse ; le pic de vitesse, par le passage par zéro de l'accélération.

Mode Naturel. Les figures 3 et 4 présentent les mesures dans le mode naturel pour AV et SL respectivement. L'enregistrement en position finale et condition très claire n'est pas disponible pour AV, et celui en position initiale et condition normale ne l'est pas pour SL. Les mesures sont en trait pointillé pour σ1 et trait plein pour σ4.

La durée des syllabes en fonction de la position de l'AP dans l'IP est donnée pour les 3 niveaux de clarté (ou de débit) dans les 3 panneaux du haut. Pour AV, dans chaque condition, et quelle que soit la position dans l'IP, la durée de σ4 est plus longue que celle de σ1. La durée des syllabes augmente avec le niveau de clarté, de même que les différences entre σ1 et σ4 s'accroissent. Pour SL, les syllabes σ4 sont en général plus longues que σ1 (une exception), la durée augmente avec la diminution du débit, mais les différences entre σ1 et σ4 ne croissent pas.

Le déplacement de la bobine située au milieu de la langue est fourni dans les panneaux du milieu. Il correspond à la différence entre une position de référence (i.e. Ymili la plus élevée atteinte pour /l/ sur l'ensemble du corpus) et Ymili pour /a/. Pour AV, le déplacement tend à être plus important pour σ4 que σ1, en positions non initiales. En position initiale, le déplacement pour σ1 peut dépasser celui pour σ4. C'est aussi pour cette position dans l'IP que l'accent LHi est le plus marqué acoustiquement. Pour SL, on observe le même phénomène, σ4 étant associée à un plus grand déplacement que σ1 en positions non initiales et le déplacement de σ1 étant supérieur à celui de σ4 en position initiale.

Le pic de vitesse figure dans les panneaux du bas. Pour SL, ses variations suivent celles du déplacement du milieu de la langue : en position non-initiale, le pic de vitesse est plus élevé pour σ4 que σ1 et l'inverse est observé en position initiale. Pour AV, la tendance est la même, avec des exceptions liées probablement à un contrôle moins clair du débit, dû à la difficulté de la tâche de clarté.

L'accent secondaire est souvent décrit comme plus faible sur le plan acoustique que l'accent primaire (Hi moins élevé que H* et durée moins longue). Mais il existe des conditions d'élocution dans lesquelles l'accent secondaire peut devenir prépondérant. Ici, il semble que σ4 soit plus articulée (augmentation de la durée, du déplacement lingual vers /a/ et du pic de vitesse de /l/ à /a/) que σ1 lorsque l'AP est en position non initiale. En position non initiale, l'articulation de σ1 (qui semble porter un accent secondaire) est donc plus faible que celle de σ4 (accent primaire). Par contre, en position initiale, l'articulation de σ1 devient prépondérante.

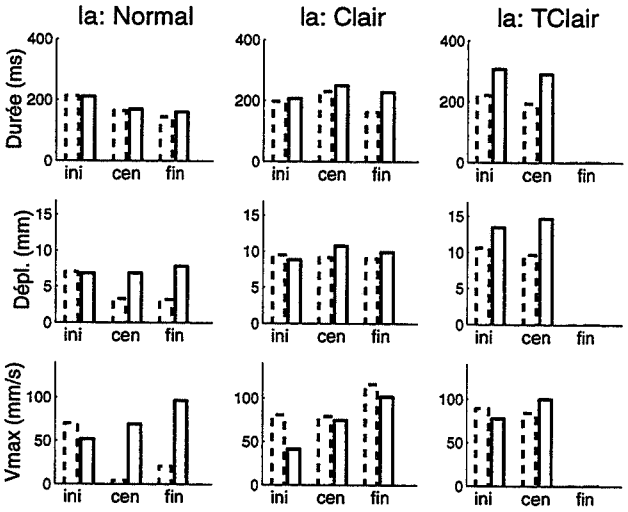


Figure 3 : Durée (ms), déplacement du milieu de la langue (mm) et pic de vitesse (mm/s) pour σ1 (pointillé) et σ4 (trait plein), mode naturel, locuteur AV.

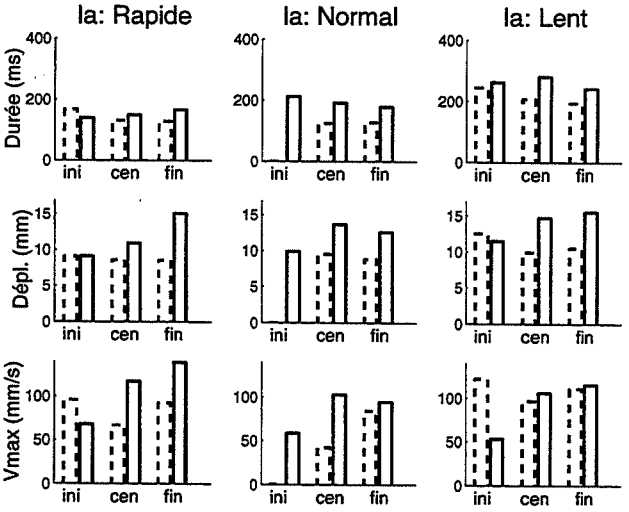


Figure 4 : idem figure 3, locuteur SL.

Mode Contrastif. Il a souvent été montré que l'emphase contrastive rend plus marqués les effets articulatoires (cf. [Jon91], [Eri98]). Les figures 5 et 6 présentent les mesures dans le mode contrastif pour SL et AV respectivement. Pour SL, l'impact de l'emphase semble réduit : la durée, le déplacement et le pic de vitesse augmentent à peine en présence d'emphase. En positions non initiales, σ4 est toujours plus fortement articulée que σ1 et la différence entre les deux syllabes tend à être plus marquée qu'en mode naturel. En position initiale, σ4 tend également à être prépondérant devant σ1 (sauf à débit rapide). Cependant, pour AV, la tendance de σ4 à être plus articulée que σ1 n'est pas renforcée en mode contrastif. Au contraire, aux débits rapide et normal, le déplacement pour σ1 est plus important que pour σ4 (quelle que soit la position dans la phrase) et la durée et le

pic de vitesse suivent (à une exception près pour chacun). A débit lent, le déplacement pour σ4 tend à être plus important que σ1, mais les différences sont ténues et les durées se comportent à l'opposé. Il semble donc que, pour AV, l'emphase contrastive ait renforcé l'accent sur σ1, devenu plus fort que l'accent sur σ4 (sauf à débit lent).

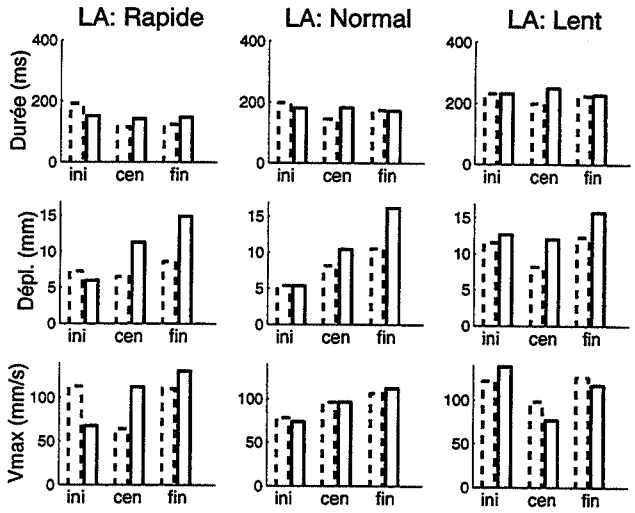


Figure 5 : Caractéristiques articulatoires de σ1 (pointillé) et σ4 (trait plein), mode contrastif, locuteur SL.

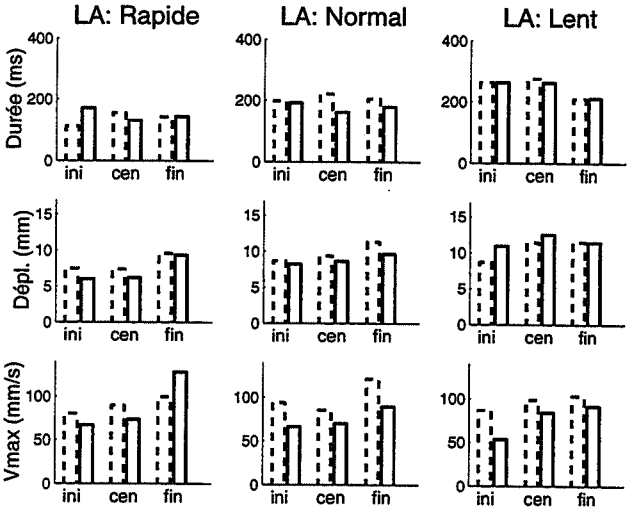


Figure 6 : idem figure 5, locuteur AV.

4. CONCLUSION

Ces résultats préliminaires laissent entrevoir les effets de l'emphase sur l'articulation de la Phrase Accentuelle en français. En mode naturel, l'AP, qui a été définie sur le plan acoustique par /LHiLH*/, semble caractérisée par un effort articulatoire plus important pour σ4 que σ1 lorsque l'AP est en position non initiale dans l'IP. Par contre, lorsque l'AP est en position initiale, σ1 devient articulatoirement (et acoustiquement) prépondérante devant σ4. L'effet de l'emphase est variable selon les locuteurs. Pour SL, l'effet de l'emphase est réduit et semble augmenter légèrement la prépondérance de σ4 devant σ1. Il semblerait donc que l'effet de l'emphase pour ce locuteur soit un léger renforcement de l'accent primaire. Cependant, l'ancrage à gauche de LHi étant assez libre, il reste à étudier les caractéristiques articulatoires de σ2 en présence d'emphase. Pour AV, l'effet de l'emphase est plus important et σ1 devient autant, voire plus, articulé que σ4. Pour cette locutrice, l'emphase induirait donc un renforcement de l'accent secondaire. L'accent dit « secondaire » peut donc devenir, sous emphase, l'accent primordial. Ce phénomène est peut-être lié à « l'accent d'insistance », apparaissant au début des mots dans certains styles de parole [Vai83].

Remerciements : À Pascal Perrier pour son aide précieuse lors du recueil des données et aux 2 patients locuteurs.

BIBLIOGRAPHIE

[Bec96] Beckman M. E. (1996). The parsing of prosody. *Language and Cognitive Processes*, 11 (1/2), 17-67.

[Eri98] Erickson D. (1998). Effects of contrastive emphasis on jaw opening. *Phonetica*, 55, 147-169.

[Fou98] Fougeron C. & Jun S.-A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *J.Phonetics*, 26, 1, 45-69.

[Hir98] Hirst D. & Di Cristo A. (1998). *Intonation systems: a survey of twenty languages.* Cambridge: Cambridge University Press.

[Jon91] de Jong K., Beckman M. E. & Fletcher J. (1991). The articulatory kinematics of final lengthening. *J. Acoust. Soc. Am.*, 89, 369-382.

[Jun95] Jun S.-A. & Fougeron C. (1995). The accentual phrase and the prosodic structure of French. *Actes du XIIIth ICPhS, Stockholm, Suède,* 2, 722-725.

[Lœv99] Lœvenbruck H. (1999). An investigation of articulatory correlates of the Accentual Phrase in French. *Proceedings of the XIVth ICPhS, San Francisco, CA, August 1999.* 1, 667-670.

[Mer93] Mertens P. (1993). Intonational grouping, boundaries and syntactic structure in French. *Proceedings of the ESCA Workshop on Prosody, Lund,* 41, 155-159.

[Ros85] Rossi M. (1985). L'intonation et l'organisation de l'énoncé. *Phonetica,* 42, 135-153.

[Vai83] Vaissière J. (1983). Language-independent prosodic features. In Prosody: Models and measurements, A. Cutler& R. Ladd (eds.), Berlin, Springer, 53-66.

[Vai92] Vaissière J. (1992). Rhythm, accentuation and final lengthening in French. In *Music, language, speech and brain,* J. Sundberg, L. Nord & R. Carlson, eds. Wenner-Gren, Int. Symposium Series, Stockholm, 59, 108-120.

[Vat93] Vatikiotis-Bateson E. & Kelso J.A.S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *J. Phonetics*, 21, 231-265.

Reconnaissance de la parole

Adaptation d'un Système de Dictée à Grand Vocabulaire en Français dédié au Domaine Radiologique

Jean-Christophe Marcadet, Claire Waast

IBM European Speech Research
IBM France Tour Descartes 2, avenue Gambetta 92400 Courbevoie, France
Mél: marcadet@fr.ibm.com - waast@fr.ibm.com

ABSTRACT

In this paper we present a large vocabulary continuous speech recognition system devoted to radiologists. Our goal is to obtain very high real time recognition rates for real users. Starting from our standard large vocabulary continuous speech recognition system we derived a very efficient system where the acoustic model, the lexicon and the linguistic model have been adapted. Recognition tests have been performed on three set of speakers: french-french radiologist students (read speech), french-canadian radiologist physicians (read speech) and french-french radiologist physicians in activity (spontaneous speech). The standard and the radiologist systems share the same architecture. Detailed comparative results will be presented on the three test sets. Roughly, each new componant introduced into the standard system individually provides 20 to 60% of relative improvement.

1. INTRODUCTION

Dans cet article, nous présentons un système de dictée, grand vocabulaire, en parole continue, dédié au domaine radiologique. Notre but est la réalisation d'un système de dictée temps réel avec un taux de reconnaissance élevé. Partant du produit standard de dictée d'IBM, nous avons adapté un système performant en reconstruisant un modèle acoustique et un modèle linguistique dédiés. Les tests de reconnaissance ont portés sur trois types de populations : des étudiants français en médecine (textes lus), des radiologues québecquois (textes lus) et des radiologues français (parole spontanée). Le système standard, ainsi que le système dédié aux radiologues, partagent la même architecture. Les résultats comparatifs sur les trois jeux de tests seront présentés. Nous verrons que chacun de ces composants apporte une amélioration significative de 20 à 60%.

2. DESCRIPTION DE L'ARCHITECTURE DU SYSTÈME

Le décodeur est constitué d'un module de traitement du signal, d'un module acoustique comportant deux passes, d'un module linguistique et d'un contrôleur. La reconnaissance est indépendante du locuteur. Un apprentissage optionnel permet d'améliorer encore les performances de reconnaissance. Le signal issu du microphone est échantillonné à une fréquence de 11kHz (16 bits). Son traitement consiste à calculer, toutes les 10 ms, 13 coefficients cepstraux sur 24 bandes de fréquences Mel en utilisant une fenêtre de 25 ms pour la FFT ainsi que les dérivées de premier et second ordre [You96]. Le système utilise des modèles acoustiques de Markov caché (ou HMM) à densités continues. Afin de prendre en compte les variations phonologiques que l'on peut observer dans différents contextes phonétiques, les modèles représentent des unités sub-phonétiques et sont dépendants du contexte phonétique. Ils sont obtenus par classification à l'aide d'arbres de décision construits à partir d'un volume important de données d'apprentissage [Bah93]. Pour chaque feuille, les paramètres acoustiques qui caractérisent les données d'apprentissage sont modélisés par une somme pondérée de densités de probabilité gaussiennes dont les matrices de covariance sont diagonales (on parle aussi de prototypes pour désigner les gaussiennes pondérées). Les modèles de Markov cachés utilisés pour modéliser les feuilles sont de simples modèles à un état avec une boucle et une transition de sortie. A partir des vecteurs acoustiques et des modèles acoustiques, on effectue un premier décodage acoustico-phonétique. Guidé par le dictionnaire de prononciations, le module acoustique extrait une liste de mots candidats. Pour chaque mot candidat, le module linguistique calcule à l'aide d'une combinaison de modèles de langage trigramme et triclasse la probabilité du mot considéré connaissant les deux mots qui le précèdent. Ce score permet de réordonner la liste des mots candidats. La seconde passe du module acoustique a un rôle similaire à celui de la première passe mais, n'ayant qu'un nombre réduit de mots à analyser, elle peut calculer plus finement l'adéquation entre vecteurs acoustiques et modèles acoustiques. Au cours de cette étape, une partie seulement des mots candidats sera retenue. Le contrôleur de décodage gère l'ensemble des trois modules précédents. Il organise les hypothèses à l'échelle de la phrase et décide de la phrase finalement décodée. L'algorithme de décodage est connu sous le nom de "décodage à pile" [Jel82][Gop95]. Un treillis de mots candidats est

construit. Pour chaque nouvelle liste d'hypothèses, la probabilité des chemins partiels est recalculée et les chemins sont triés par ordre de probabilité décroissante. Seuls les chemins les plus probables sont conservés et étendus. Le modèle acoustique peut être utilisé tel quel en mode indépendant du locuteur ou peut être adapté au locuteur par un apprentissage supervisé, pour une précision accrue. Cet apprentissage consiste à faire prononcer au locuteur test un certain nombre de phrases préétablies. Une estimation Bayesienne [Gau94] est utilisée pour réestimer les prototypes initiaux en fonction des nouvelles données. Seuls les prototypes qui changent de façon significative sont adaptés.

3. UN SYSTÈME DÉDIÉ AUX RADIOLOGUES

3.1 Le système standard

En ce qui concerne les modèles acoustiques, le système repose sur un ensemble de 38 phonèmes. Chaque phonème est découpé en trois unités séquentielles sub-phonétiques que l'on peut interpréter comme l'attaque, le milieu et la chute du phonème. A chaque unité sub-phonétique est associé, à l'aide d'arbres de décision, des classes (ou feuille) dépendant du contexte phonétique. A chaque feuille est associée une somme pondérée de gaussiennes. Dans le souci de limiter les phénomènes de sous-estimation des paramètres statistiques ainsi que le temps de calcul au décodage, le nombre des feuilles ainsi que celui des gaussiennes est limité à typiquement 3000 feuilles et 30 000 prototypes. Différentes méthodes peuvent être utilisées pour obtenir ces prototypes [Che98][Bah98]. Celle retenue pour le système standard ainsi que pour le système de radiologie impose au nombre des prototypes de chaque feuille d'être proportionnel au nombre d'échantillons d'apprentissage disponibles [Bah95]. Les données d'apprentissage consistent en un ensemble de 160 000 phrases lues (320 heures) par 900 locuteurs équitablement distribués en hommes et femmes. Les enregistrements ont été réalisés en milieu sonore calme (bureau ou équivalent) avec différents types de microphones. Les locuteurs ont pour une petite moitié un accent parisien, puis également répartis entre les catégories "Sud", "Nord", "Est", "Ouest" de la France et "autre" correspondant en majorité à des locuteurs d'origine étrangère. En âge, 2/3 des locuteurs ont entre 25 et 40 ans, 1/6 ont moins de 25 ans et 1/6 ont plus de 40 ans. Tous les textes lus sont différents et sont issus en grande majorité de textes journalistiques. Les enregistrements ayant été menés sous la surveillance d'un superviseur, ils n'ont pas donné lieu à des transcriptions. Leur validité est jugée à travers un alignement de Viterbi qui conduit à un rejet de moins de 5% des phrases. Le dictionnaire de prononciations utilisé en phase d'apprentissage ainsi que celui utilisé en phase de décodage sont obtenus par une méthode mixte. IBM a depuis plusieurs années développé une base de données multi-indices qui à chaque forme fléchie associe des informations lemmatiques, syntaxiques et phonétiques. Si un mot du dictionnaire de prononciations est connu de cette base, ses formes phonétiques en sont issues. Sinon les formes phonétiques du mot sont calculées par un phonétiseur par règle avec variantes [Waa91]. Ces formes sont validées par relecture puis intégrées dans la base de données. Les textes d'apprentissage sont formatés de sorte à être en adéquation parfaite avec la notion de "mots" induit par le modèle de langage. En particulier les mots composant les locutions seront appris acoustiquement comme appartenant à un même groupe. La réalisation des liaisons est toujours optionnelle. Elle est prédite en fonction d'informations contenues dans les dictionnaires [Waa98]. Enfin l'optimisation des paramètres de décodage est faite de sorte à obtenir une vitesse de décodage de l'ordre de 80% du temps de dictée sur un processeur Intel Pentium 200MHz.

Le modèle de langage de référence est une combinaison linéaire d'un modèle trigramme et d'un modèle triclasse [Jst97]. Ces modèles ont été collectés sur des comptes-rendus issus d'hôpitaux et cabinets privés, soit un total de plus de 50 millions de mots. Le vocabulaire issu de ces textes comporte les 50.000 mots les plus fréquents, soit un nombre de prononciations total de 63.000.

3.2 Un modèle acoustique pour les radiologues

Le modèle acoustique pour les radiologues est obtenu à partir du modèle standard, en itérant deux fois le processus de construction des arbres phonologiques et du calcul des prototypes. Les données d'apprentissage consistent en un ensemble de 110 000 phrases (315 heures) par 580 locuteurs répartis en 2/3 hommes, 1/3 femmes. Parmi ces locuteurs, 200 sont des étudiants en médecine qui ont lu des comptes-rendus d'examen radiologique, 300 sont des radiologues qui pour moitié ont lu leurs comptes-rendus et pour moitié les ont composés au moment de l'enregistrement, 30 sont des radiologues québecquois qui ont lu leurs comptes-rendus, 50 sont des québecquois ayant lu des textes sans rapport à la radiologie. Pour les locuteurs français, les distributions en âge et en accent sont comparables à celles précédemment exposées. Les comptes-rendus spontanés des radiologues ont été transcrits en annotant les hésitations, toux, rires etc.

3.3 Un modèle linguistique pour les radiologues

Le modèle linguistique dédié au système de reconnaissance pour la radiologie est obtenu à partir du modèle de langage du système de référence en

ajoutant de nouveaux mots issus de nouveaux corpora. Ces derniers proviennent de deux grands hôpitaux français et d'un hôpital québecquois. Cette collecte complémentaire a permis de porter le total de mots pour la modélisation à plus de 83 millions de mots. Tout comme le modèle linguistique du système de référence, nous avons combiné linéairement deux modèles trigramme et triclasse grammaticale. L'étude minutieuse des textes d'apprentissage nous a permis d'extraire un vocabulaire de 52.000 mots, 65.000 prononciations, intégrant de nombreuses locutions et abréviations propres à la profession. Une attention particulière a été portée à la phonétisation des termes médicaux avec une relecture manuelle et une vérification par un radiologue professionnel. Nous avons également actualisé la liste des noms propres avec des mots issus des dernières technologies d'investigation, d'appareillage, de produits de contraste et médicaments. D'autre part, nous avons tenu compte des accents régionaux français ainsi que québecquois. Afin de mieux prendre en compte la dictée spontanée, quelques mots modélisant les hésitations ont été ajoutés.

3.4 Un texte d'apprentissage pour les radiologues

L'adaptation supervisée du modèle acoustique nécessite la création d'un texte phonétiquement équilibré (texte d'apprentissage). Nous avons sélectionné 150 phrases (environ 53 mille mots), issues de comptes-rendus de radiologie. Chaque phonème y est représenté un minimum de 90 fois et dans le plus grand nombre de contextes possibles. Le choix d'utiliser des phrases médicales pour ce texte d'apprentissage s'est avéré un point d'ergonomie important pour les utilisateurs finaux.

4. RÉSULTATS

L'évaluation des performances consiste à mesurer, en comparaison avec le système standard, les gains résultant d'une part du nouveau modèle acoustique, d'autre part du nouveau modèle linguistique et enfin de l'adaptation au locuteur. Les données de test se divisent en trois parties : 10 locuteurs étudiant en médecine ayant lu un ensemble de comptes-rendus de 100 phrases (1973 mots) ; 10 radiologues ayant composé spontanément un ensemble de comptes-rendus d'en moyenne 94 phrases (1478 mots) et 6 radiologues québecquois ayant lu le même test que les étudiants. Les données d'adaptation au locuteur consiste en 150 phrases de radiologie pour chaque locuteur de test. Dans un premier temps, nous avons substitué le modèle acoustique des radiologues au modèle acoustique standard. Le modèle de langage dans cette expérience restant celui du système standard. Le système résultant est appelé +MA. Nous avons ensuite substitué dans le système +MA, le nouveau modèle de langage des radiologues au modèle de langage du système standard. Le système résultant est appelé +ML. Ce dernier système contient donc les nouveaux modèles acoustique et linguistique. Enfin, à partir du système +ML, nous avons procédé à l'adaptation des locuteurs (+Adapt.). Les résultats de ces expériences sont donnés dans les tableaux (table 1), (table 2) et (table 3). On peut constater que les améliorations obtenues au cours de chaque expérience sont très significatives.

Table 1: Impact du nouveau modèle acoustique

	Standard	+MA	Gain
Etudiants	9,78	7,25	26%
Radiologues	15,92	10,28	35%
Québecquois	26,11	14,50	44%

Le gain résultant de +MA tient essentiellement à la nature des données d'apprentissage. Ce modèle acoustique a été construit à partir d'enregistrements lus et spontanés d'une grande variabilité de vitesse d'élocution (de 45 à 140 mots par minutes). Ces enregistrements contiennent de nombreuses plages de silence et d'hésitations. Les données d'apprentissage utilisées pour construire le système standard sont des données lues, bien moins riches en style d'élocution. La présence d'enregistrements de locuteurs québecquois (bien qu'en nombre réduit) permet d'accroître leurs performances, sans dégradation significative de celles des locuteurs français. Le gain sur les locuteurs étudiants (textes lus) peut s'expliquer par une meilleure prise en compte de la variabilité phonétique dans l'arbre phonologique : mots anglais, beaucoup de nombres et de nombreux mots à racine latine ou grecque.

Table2: Impact du nouveau modèle linguistique

	+MA	+ML	Gain
Etudiants	7,25	5,46	25%
Radiologues	10,28	8,46	18%
Québecquois	14,50	12,13	16%

L'ajout de 32 millions de mots dans le corpus d'apprentissage permet d'améliorer les statistiques des modèles de langage et de diminuer le nombre de mots inconnus (2 mots inconnus avec le vocabulaire standard et 0 avec le nouveau vocabulaire). Le gain observé dans le tableau précédent est essentiellement dû à d'une part l'amélioration des formes phonétiques des mots du vocabulaire et à l'introduction de mots modélisant les hésitations et d'autre part à l'ajout d'abréviations. L'apport de textes en provenance d'un hôpital québecquois n'a pas montré un gain significatif sur les locuteurs canadiens ; ceci s'explique par le fait que les comptes-rendus radiologiques sont très similaires à ceux en provenance des hôpitaux français; la richesse du vocabulaire québecquois et les expressions typiques de nos cousins canadiens n'apparaissent pas dans ce type de textes techniques...

Table3: Impact de l'adaptation aux locuteurs de test

	+ML	+Adapt.	Gain
Etudiants	5,46	3,67	33%
Radiologues	8,46	6,32	25%
Québecquois	12,13	4,83	60%

Dans des conditions similaires le gain résultant de l'adaptation est, pour le système standard, de l'ordre de 25% sur de la parole lue. En comparaison, le système dédié aux radiologues apporte un gain légèrement supérieur (33% sur les étudiants). Le gain plus modéré sur les radiologues peut s'expliquer par le fait que l'adaptation aux locuteurs s'est faite sur des phrases lues, alors que le test de décodage s'est déroulé sur des phrases spontanées. On notera la très forte adaptation acoustique aux locuteurs québecquois ; ceci est imputable à la faible proportion d'enregistrements canadiens la base de données de construction du modèle acoustique. Le taux d'erreur en mode indépendant du locuteur sur les Canadiens est de 12,13%, la technique d'adaptation aux locuteurs ramène ce taux d'erreur à un taux proche de celui des étudiants français.

5. CONCLUSION

Nous avons montré que l'adaptation d'un système de dictée grand vocabulaire en français à un domaine technique spécialisé comme la radiologie nécessite l'adaptation du vocabulaire et du modèle de langage, mais aussi celle du modèle acoustique. L'intégration de données réelles et proches de l'utilisation finale du système, comme l'ajout d'un très grand nombre de comptes-rendus médicaux, l'enregistrement de textes en dictée spontanée et d'enregistrements de locuteurs correspondant à la population visée, permet d'obtenir un système global de très hautes performances.

BIBLIOGRAPHIE

[Jel82] Jelinek F., Mercer R., Bahl L.R. (1982) "Continuous speech recognition : statistical methods." Handbook of statistics, Vol.2. Classification, pattern recognition and reduction of dimensionality. P. Krishnaiah, L Kanal, North Holland.

[Waa91] Waast C. (1991) "Phonétiseur du français avec variantes: son intégration dans un système de reconnaissance probabiliste" AFCET RFIA

[Bah93] Bahl L.R., de Souza P., Gopalakrishnan P., Picheny M. (1993) "Context-dependent vector quantization for continuous speech recognition" ICASSP

[Gau94] Gauvain J.L., Lee C.H (1994) "Maximum a-posteriori estimation for multivariate gaussian mixture observations of Markov chain" IEEE Transactions on Speech and Audio Proc. Vol 2. pp 291-298

[Bah95] Bahl L.R. et al. (1995) "Performance of the IBM large vocabulary continuous speech recognizer on the ARPA Wall Street Journal task" ICASSP, pp. 41-44.

[Gop95] Gopalakrishnan P.S., Bahl L.R., Mercer R. (1995) "A tree search strategy for large-vocabulary continuous speech recognition" ICASSP

[You96] Young S. (1996) "A review of large-vocabulary continuous speech recognition." IEEE Signal Proc. magazine, Sept. 1996.

[Jst97] Crépy H., Marcadet J.C., Waast C.(1997) "Dictée à Grand Vocabulaire en Français : IBM VoiceType 3.0, un produit de la Recherche" JST 97 Avignon pp 19-23

[Bah98] Bahl L.R. , Padmanabhan M. (1998) "A discriminant measure for model complexity adaptation" ICASSP, pp. 453-456

[Che98] Chen S.S., Gopalakrishnan P.S. (1998) "Clustering via the Bayesian Information Criterion" ICASSP, pp. 645-648

[Waa98] Bahl L.R. et al. (1998) "A method for modeling liaison in a speech recognition system for French" ICSLP pp 114-117

Détermination d'une Mesure de Confiance pour le Rejet des Entrées Incorrectes

Nicolas Moreau, Denis Jouvet

France Télécom – CNET / DIH / DIPS
2, avenue Pierre Marzin, 22307 Lannion Cédex
Tél.: +33 (0)2 96 05 31 52 - Fax: +33 (0)2 96 05 35 30
Mél: nicolas.moreau@cnet.francetelecom.fr

ABSTRACT

Interactive vocal services are based on speech recognition systems which must be able to reject efficiently incorrect utterances such as out-of-vocabulary or noise tokens. A possible approach is a post-processing of the hypotheses delivered by the recogniser, based on the computation of a confidence measure (CM). A recognition hypothesis is rejected if its CM is below a chosen threshold. This paper presents and compares several ways of computing a CM on a recognition hypothesis, based on the calculation of a likelihood ratio for each acoustic frame of the utterance. Results are reported on a large vocabulary of a telephone directory task. A significant decrease in the error rates is observed, compared to a reference system which includes only a garbage model, with no post-processing of the recognised words.

1. INTRODUCTION

Les applications pratiques de la reconnaissance automatique de la parole sont souvent confrontées à des utilisateurs peu conscients des contraintes du système et s'exprimant dans un environnement sonore bruité, comme c'est le cas, par exemple, pour les services vocaux interactifs sur le réseau téléphonique.

Les systèmes de reconnaissance doivent alors être capables de rejeter les entrées incorrectes que sont les mots ou phrases ne faisant pas partie du vocabulaire de l'application ainsi que les diverses perturbations accidentelles (hésitations de l'utilisateur, bruit environnant, etc.). De tels systèmes sont confrontés à trois grands types d'erreurs :

- erreurs de *rejet à tort* d'expressions valides.
- erreurs de *substitution* (un mot du vocabulaire de l'application est reconnu à la place d'un autre).
- erreurs de *fausse alarme* (une entrée incorrecte est prise pour un mot ou une phrase du vocabulaire).

Les conséquences de ces erreurs sont diverses. En général, les systèmes existants conduisent à un taux d'erreur de substitution raisonnablement bas. Les erreurs de rejet à tort ne sont pas gênantes si elles ne sont pas trop nombreuses : elles peuvent donner lieu, dans le cas d'un service vocal interactif par exemple, à une demande de répétition de la part du système. Les erreurs de fausse alarme, enfin, sont les plus pénalisantes puisqu'elles entraînent des actions qui n'ont pas été demandées par l'utilisateur.

Dans ce contexte, il peut être très profitable de calculer une mesure de confiance pour chaque unité reconnue par le système (phonème, mot, etc.), c'est-à-dire d'associer à chacune de ces hypothèses de reconnaissance une mesure traduisant sa fiabilité. Cette information peut ensuite être comparée à un seuil pour permettre le rejet des hypothèses de reconnaissance les moins fiables.

Diverses méthodes ont été proposées pour calculer une mesure de confiance sur une hypothèse de reconnaissance W. On peut distinguer deux approches principales du problème. La première est basée sur l'estimation de la probabilité que W soit correcte. C'est le cas notamment des approches par arbres de décision [Cha97] ou par modèles linéaires généralisés [Gil97]. La seconde, qui est le fondement de cette étude, repose sur la théorie des tests d'hypothèse, théorie largement utilisée dans le domaine de la vérification du locuteur et de l'information verbale [Lee97]. La mesure de confiance sur l'hypothèse W est dans ce cas assimilée au rapport de deux probabilités : la probabilité que W soit correcte sur la probabilité que W soit incorrecte. Là encore, un grand nombre de solutions ont été proposées pour faire l'estimation d'un tel rapport.

Dans le cas d'un système de reconnaissance à base de modèles de Markov, l'approche la plus classique consiste à prendre le rapport de deux scores acoustiques [Rah97]. Le numérateur est alors le score résultant de l'alignement sur le modèle markovien de W. Le dénominateur est soit la combinaison des scores d'une *cohorte* de modèles compétiteurs, soit le score d'un *anti-modèle* associé à W.

Dans [Bar97] et [Jou99] on trouve cependant une méthode mieux adaptée au cas d'une modélisation flexible (par phonèmes). Un modèle et un anti-modèle sont appris pour chaque phonème φ, à partir de paramètres segmentaux (phonétiques et prosodiques). On peut alors calculer un rapport de vraisemblance sur toute portion de signal alignée sur φ. L'idée est de combiner les résultats obtenus pour chaque phonème reconnu afin d'obtenir un rapport de vraisemblance global.

Cet article repose sur une méthode – introduite dans [Mor99] – qui s'inspire du même principe, et analyse certains aspects de la modélisation mise en œuvre. La mesure de confiance proposée est la combinaison de rapports de vraisemblance calculés au niveau le plus élémentaire : celui des trames acoustiques.

2. MESURE DE CONFIANCE

Dans cette étude, les mesures de confiance sont utilisées pour décider du rejet ou de l'acceptation d'une hypothèse de reconnaissance. La prise de décision s'appuie sur un test du rapport de vraisemblance [Lee97].

2.1 Test du rapport de vraisemblance

Le test s'écrit de la manière suivante, en notant W le résultat du décodage du signal d'entrée X :

$$LR(X \mid W) = \frac{P(X \mid W \text{ correct})}{P(X \mid W \text{ incorrect})} \begin{cases} \geq \omega \Rightarrow W \text{ acceptée} \\ < \omega \Rightarrow W \text{ rejetée} \end{cases} \quad (1)$$

Toute hypothèse W pour laquelle le rapport $LR(X \mid W)$ est inférieur à un seuil ω qu'on s'est fixé, est rejetée.

2.2 Calcul au niveau de la trame

Nous proposons ici d'estimer le rapport $LR(X \mid W)$ en combinant des rapports de vraisemblance calculés au niveau des trames acoustiques.

Après alignement de l'entrée X sur le modèle de Markov associé à W, chaque trame x de X est associée à un état acoustique q. On définit le rapport de vraisemblance sur la trame x de la façon suivante :

$$LR(x \mid q) = \frac{P(x \mid M_q)}{P(x \mid M_{\overline{q}})} \quad (2)$$

On note $P(x \mid M_q)$ et $P(x \mid M_{\overline{q}})$ les scores de x sur deux modèles M_q et $M_{\overline{q}}$. Le premier, M_q, est le modèle des événements corrects associé à l'état q. Le second, $M_{\overline{q}}$, est le modèle des événements incorrects, ou *anti-modèle,* associé à ce même état q. La nature de ces modèles sera décrite dans la partie suivante.

Le rapport de vraisemblance $LR(X \mid W)$ associé à X s'obtient en combinant les rapports de vraisemblance calculés sur chacune des trames de X. Si on note $Q = (q_1, q_2 \ldots, q_T)$ la séquence d'états acoustiques correspondant au décodage de $X = (x_1, x_2 \ldots, x_T)$, le rapport de vraisemblance global s'écrit :

$$LR(X \mid W) = \frac{\prod_{i=1}^{T} P(x_i \mid M_{q_i})}{\prod_{i=1}^{T} P(x_i \mid M_{\overline{q_i}})} \quad (3)$$

Au final, on obtient la mesure de confiance $CM(W)$ sur l'hypothèse W en normalisant le logarithme de ce rapport par le nombre T de trames acoustiques de l'entrée X :

$$CM(W) = \frac{1}{T} Log\,[LR(X \mid W)] \quad (4)$$

3. MODELISATION

Cette approche nécessite l'apprentissage d'un modèle M_q et d'un anti-modèle $M_{\overline{q}}$ pour chaque état q des modèles de Markov du décodeur. La principale difficulté réside dans la définition de $M_{\overline{q}}$, qui doit modéliser différents types d'événements incorrects.

3.1 Anti-modèles

L'anti-modèle $M_{\overline{q}}$ associé à q est ici constitué d'un ensemble de densités de probabilité, qui modélisent chacune un type d'erreur différent. Trois densités ont été apprises pour chaque état q : $M_{\overline{q}(sub)}$ pour les erreurs de substitution, $M_{\overline{q}(hv)}$ pour les erreurs de fausse alarme sur les mots hors vocabulaire et $M_{\overline{q}(br)}$ pour les erreurs de fausse alarme sur les bruits. Ces densités sont estimées à partir des trames acoustiques associées à q au sein d'alignements incorrects (substitutions, fausses alarmes sur des mots hors vocabulaire ou des bruits, respectivement).

Le score $P(x \mid M_{\overline{q}})$ d'une trame x sur l'anti-modèle $M_{\overline{q}}$ s'obtient en combinant les vraisemblances de x sur ces différentes densités. Plusieurs fonctions de combinaison ont été testées ([Mor99]). Nous faisons ici le choix de la moyenne des vraisemblances :

$$P(x \mid M_{\overline{q}}) = \frac{1}{3}\left[M_{\overline{q}(sub)}(x) + M_{\overline{q}(hv)}(x) + M_{\overline{q}(br)}(x)\right] \quad (5)$$

Il est possible de ne combiner que deux des trois vraisemblances précédentes. Dans ce qui suit, la moyenne des deux vraisemblances $M_{\overline{q}(hv)}(x)$ et $M_{\overline{q}(br)}(x)$ sera également utilisée.

Enfin, le modèle M_q est une densité de probabilité apprise, de la même manière que les densités précédentes, à partir des trames associées à q dans un corpus d'alignements corrects.

3.2 Estimation des densités

Toutes les densités de probabilité sont estimées dans le même espace de paramètres acoustiques que celui de la modélisation markovienne (il s'agit de coefficients MFCC et de leurs dérivées premières et secondes). L'un des intérêts de cette approche est qu'elle ne nécessite l'extraction d'aucun paramètre de post-traitement supplémentaire.

Deux méthodes d'estimation ont été employées. La première fait l'hypothèse de densités gaussiennes continues. La seconde consiste à estimer un ensemble de densités discrètes. Elles sont obtenues à partir

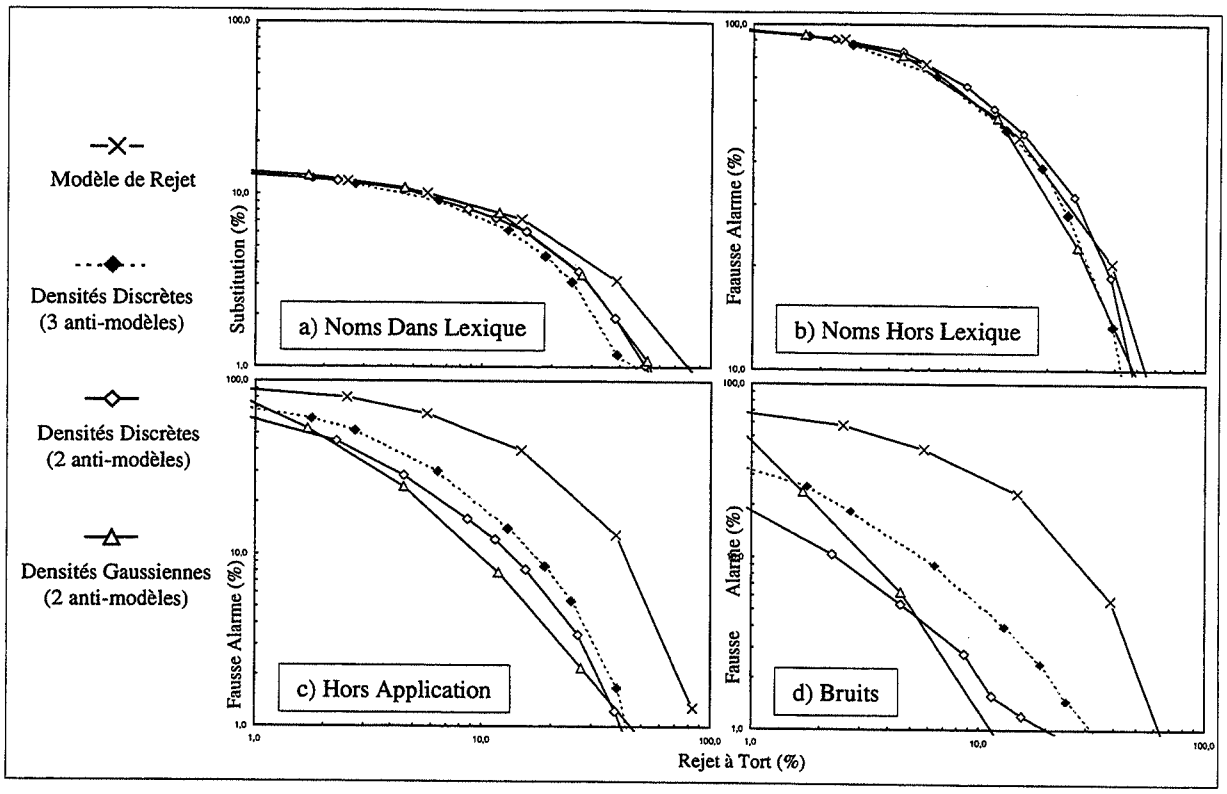


Figure 1: Résultats de post-traitement pour différents modèles.

d'histogrammes de données dans l'espace acoustique discrétisé (trois sous-espaces discrétisés sur 2^6 points chacun).

4. EXPERIENCES

Une procédure de rejet des entrées incorrectes, basée sur la mesure de confiance précédemment décrite, a été évaluée sur un annuaire vocal à grand vocabulaire.

4.1 Annuaire vocal

Les évaluations ont été réalisées sur une base de données d'exploitation d'un annuaire vocal. La consultation téléphonique de cet annuaire se fait en donnant le nom, éventuellement précédé du prénom, de la personne à joindre. Le vocabulaire de l'application comporte 1587 mots. Les corpus d'apprentissage C_{app} (15244 entrées) et de test C_{test} (6353 entrées) sont composés d'entrées correctes (a) (50% de C_{test}), de bruits (d) (15%), et de données hors vocabulaire. Pour ces dernières, la distinction a été faite entre les données non valides mais proches de l'application (prénom et nom) (b) (10%), et les autres (c) (25%). Le sous-corpus (b) rassemble les entrées où le prénom et/ou le nom ne sont pas répertoriés dans le lexique, ainsi que les séquences prénom plus nom qui ne sont pas autorisées. Les données (c) résultent d'une utilisation incorrecte de l'application (hésitations, interruptions, phrases non conformes au mode d'emploi de l'annuaire).
Le système repose sur une modélisation markovienne flexible (par modèles de phonèmes en contexte) des noms et prénoms répertoriés dans l'annuaire.

4.2 Performances sur les différents corpus

Les courbes de la figure 1 sont obtenues en faisant varier le seuil de décision du test d'hypothèse (Eq.1). Elles donnent l'évolution du taux de substitution (a) et des taux de fausse alarme sur les noms hors lexique (b), les données hors application (c) et les bruits (d), en fonction du taux de rejet à tort.

Les performances de notre procédure de rejet à base de CM sont comparées à celles d'un système de référence (×) se servant d'un modèle générique de rejet, sans aucun post-traitement, pour écarter les entrées incorrectes. La courbe (×) est obtenue en faisant varier le coût associé au modèle de rejet.

Les trois autres courbes de la figure 1 sont obtenues par post-traitement des hypothèses en prenant la moyenne arithmétique (Eq.5) des scores d'anti-modèles (d'autres méthodes de combinaison ont été testées dans [Mor99]). Par rapport au système de référence, on ne constate pas d'amélioration significative du taux de substitution (a), qui reste autour de 8% pour un taux de rejet à tort de 10%. On observe en revanche une amélioration très nette des performances de rejet sur les données hors application (c) et les bruits (d). Pour un taux de rejet à tort de 10%, on obtient – courbe (Δ) – des réductions relatives des taux de fausse alarme allant jusqu'à 80% (de 50% de fausses alarmes avec modèle de rejet seul à 10% après post-traitement) sur le corpus (c) et de plus de 90% (de 30% à 1,2%) sur les bruits (d). Cependant, dans tous les cas de figure, le post-traitement ne s'avère pas plus efficace que le système

de référence pour rejeter les noms et prénoms hors lexique (b). Le taux de fausse alarme stagne autour de 60% (pour un rejet à tort de 10%) quelle que soit la méthode de rejet envisagée.

4.3 Influence du nombre d'anti-modèles

Pour les densités discrètes, deux jeux d'anti-modèles ont été utilisés. Dans le premier cas (◆) on utilise trois anti-modèles ($M_{\overline{q}(sub)}$, $M_{\overline{q}(hv)}$ et $M_{\overline{q}(br)}$). Dans le second (◇), on n'en conserve que deux ($M_{\overline{q}(hv)}$ et $M_{\overline{q}(br)}$).

Les performances de rejet sur les corpus (c) et (d) s'améliorent si on laisse de côté le modèle $M_{\overline{q}(sub)}$ appris sur les erreurs de substitution (respectivement réduction de 30% et 60% des fausses alarmes pour un rejet à tort de 10%). Cela se fait au prix d'une dégradation logique du taux de substitution (on observe le même phénomène avec des densités gaussiennes). Mais cette dégradation est peu significative (le taux reste entre 7 et 8% pour un rejet à tort de 10%) ce qui justifie le choix de ne garder que deux anti-modèles pour l'expérience de la section suivante.

4.4 Influence du type de densité

L'utilisation de densités gaussiennes (Δ) a également été testée dans le cas de 2 anti-modèles. Les résultats observés sont sensiblement meilleurs que ceux de tests antérieurs [Mor99]. Cette amélioration a été obtenue en imposant une valeur minimum aux écarts types des densités (lissage des « pics » dus à un manque de données d'apprentissage pour certaines densités).

Les performances de rejet sur les données hors application sont meilleures dans le cas gaussien (Δ) que dans le cas discret (◇) (à partir d'un taux de rejet à tort de 2%). En ce qui concerne le rejet des bruits, les densités discrètes donnent de meilleurs résultats en deçà d'un taux de rejet à tort situé autour de 5%. Il est donc difficile de donner une préférence à l'une ou l'autre approche. D'autres tests ont été réalisés avec des densités multigaussiennes (mélanges de 2 ou de 4 gaussiennes), sans qu'on obtienne de meilleurs résultats qu'avec les gaussiennes simples.

5. CONCLUSION

Cet article présente une stratégie de rejet des entrées incorrectes d'un annuaire vocal par post-traitement des hypothèses de reconnaissance. Les performances de cette approche, basée sur le calcul d'une mesure de confiance, ont été comparées à celles d'un système utilisant un modèle de rejet. L'utilisation de densités gaussiennes et de deux anti-modèles semblent donner les meilleurs résultats, abaissant fortement les taux de fausse alarme sur les bruits et les entrées hors application par rapport au système de référence utilisant un modèle générique de rejet. Notons que ces deux approches ne sont pas incompatibles. L'étude [Mor00] montre qu'il est possible de combiner efficacement l'utilisation d'un modèle de rejet et d'une procédure de post-traitement.

Enfin, ce travail fait apparaître la difficulté de rejeter efficacement les entrées incorrectes qui sont proches du vocabulaire et de la syntaxe de l'application. Aucune des solutions proposées ne permet d'écarter plus de 40% des noms hors lexique. Ces données présentent souvent une importante partie commune (le prénom, par exemple) avec une entrée valide, ce qui peut amener à des mesures de confiance globales (moyenne sur l'ensemble des mots ou de l'expression) relativement élevées. Une des solutions envisageables pour contourner le problème serait de calculer des mesures de confiances distinctes sur différentes parties du signal. On pourrait ainsi détecter les portions les moins fiables et rejeter les entrées dont certains segments auraient des mesures de confiance trop faibles.

BIBLIOGRAPHIE

[Cha97] Chase L. (1997), "Confidence Annotation for Automatic Recognition of Conversational Telephone Speech", *COST 250*, pp. 33-36.

[Gil97] Gillick L., Ito Y., Young J. (1997), "A Probabilistic Approach to Confidence Estimation and Evaluation", *ICASSP'97*, vol. 2, pp. 879-882.

[Lee97] Lee C.-H. (1997), "A Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification", *COST 250*, pp. 63-72.

[Rah97] Rahim M. G., Lee C.-H., Juang B.-H. (1997), "Discriminative Utterance Verification for Connected Digit Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277.

[Bar97] Bartkova K., Jouvet D. (1997), "Usefulness of Phonetic Parameters in a Rejection Procedure of an HMM Based Speech Recognition System", *Eurospeech'97*, vol. 1, pp. 267-270.

[Jou99] Jouvet D., Bartkova K., Mercier G. (1999), "Hypothesis Dependent Thresholding Setting for Improved Out-of-Vocabulary Data Rejection", *ICASSP'99*, vol. 2.

[Mor99] Moreau N., Jouvet D. (1999), "Use of a Confidence Measure Based on Frame Level Likelihood Ratios for the Rejection of Incorrect Data", *Eurospeech'99*, vol.1, pp.291-294.

[Mor00] Moreau N., Charlet D., Jouvet D. (2000), "Confidence Measure and Incremental Adaptation for the Rejection of Incorrect Data", à paraître dans *ICASSP'2000*.

Partitionnement dynamique des distributions pour le calcul des émissions dans un décodeur acoustico-phonétique Markovien

Georges Linares, Pascal Nocera, Driss Matrouf

Laboratoire d'Informatique d'Avignon -CERI- 439 chemin des Meinajariès - 84190 Avignon CEDEX 9
Tél.: ++33 (0) 04 90 84 35 20 - Fax: ++33 (0) 04 90 84 35 01
Mél: georges.linares,driss.matrouf,pascal.nocera@lia.univ-avignon.fr

Abstract

Acoustic models training for an automatic speech recognition system based on gaussian mixtures requires important CPU ressources; the major part of consumed time is taken by the computation of the likelihood of frames given each gaussian. In this paper, we describe a method which performs significant time reduction by clustering gaussians and associating dynamically a subset of clusters to each observation vector. This dynamic partitionning of distributions allows selection of significative emissions which must be computed. Experiments on the TIMIT database show significant computing time reduction without phone error rate increase.

1. Introduction

Ces dernières années, l'essentiel des recherches en reconnaissance de la parole a porté sur l'amélioration des taux de reconnaissance sur des tâches de plus en plus complexes (environnement bruité, très grand vocabulaire, etc.). Malgré l'augmentation de la puissance des machines utilisées, la très forte consommation des ressources CPU des systèmes "état de l'art" limite leur utilisabilité dans des contextes réels, alourdit le processus de mise au point et l'évaluation d'éventuelles approches originales. Des méthodes de réduction de la complexité du décodage ont été développées et évaluées avec succès, au niveau des algorithmes de recherche de chemins optimaux dans des treillis de mots ([8]), ou du calcul des probabilités d'émission des modèles acoustiques. Les contributions de chacun de ces processus dans la durée totale de décodage sont assez difficiles à évaluer ; elles dépendent des stratégies de décodage retenues par chaque système en particulier et de la tâche traitée (Décodage Acoustico-Phonétique, grand vocabulaire, etc..). Néanmoins, Bocchieri [1] estime à plus de 95% la part du calcul des émissions dans un système petit vocabulaire, et Knill [3] estime cette contribution dans un intervalle de 30% à 70% pour un système grand vocabulaire. Nous décrivons, dans cet article, un algorithme visant à réduire la part du calcul des émisssions dans le processus d'apprentissage et de décodage.

2. Modélisation acoustique par modèles de Markov cachés

La plupart des systèmes de reconnaissance automatique de la parole (SRAP) modernes reposent sur une modélisation acoustique par des Modèles de Markov Cachés (MMC). Les unités acoustiques sont représentées par des machines de Bakis à n états ; à chaque état, est associée une fonction de densité de probabilité (FDP) qui permet le calcul de la vraisemblance d'une observation acoustique sachant l'état considéré. Généralement, les distributions sont approchées par des mixtures de gaussiennes. Les distributions réelles étant multi-modales, les mixtures doivent comporter de nombreuses composantes pour en obtenir une bonne approximation. Cette approche pose de nombreux problèmes relatifs à la qualité de l'approximation obtenue et à la complexité des modèles ; ces points ont donné lieu, ces dernières années, au développement de méthodes destinées à améliorer l'estimation des mixtures ou basées sur des approches fondamentalement différentes ([9][6]). Néanmoins, les résultats obtenus et la facilité d'intégration des modèles acoustiques markoviens à un système de dictée vocale complet font que les modèles à base de mixtures de gaussiennes restent les plus fréquemment utilisés.

L'estimation des mixtures de gaussiennes est éffectuée par itération d'un processus de segmentation puis re-estimation jusqu'à stabilisation des distributions obtenues. Si des modèles acoustiques ont déjà été estimés, la segmentation est réalisée par un décodage contraint ; le décodeur cherche dans ce cas la segmentation optimale en états connaissant la chaine phonétique de référence. Sinon, un segmentation uniforme peut-être utilisée comme segmentation intiale. L'estimation des mixtures de gaussiennes est effectuée par des algorithmes classiques de maximisation de la vraisemblance des observations connaissant le modèle associé, du type *Expectation Maximisation* (EM).

Le décodage acoustico-phonétique d'un segment de parole est réalisé par alignement dynamique d'une séquence de MMC et de la séquence d'observations, généralement par l'algorithme de Viterbi. La complexité de l'algorithme est d'ordre $n^2.t$, où n est le nombre total d'états des modèles, et t le nombre de trames de la séquence à décoder. Bien plus que la procédure d'alignement dynamique elle-même, c'est le calcul des vraisemblances qui pénalise la vitesse

de décodage. Ainsi, pour un système simple comportant 48 modèles à 3 états, et avec des mixtures de 8 gaussiennes, c'est environ 10^3 vraisemblances par trame qui doivent être calculées. Cependant, les scores des distributions modélisant des régions de l'espace acoustique éloignées de l'observation ne sont pas significatives, et leur calcul ne fait qu'alourdir le processus de décodage. L'objectif des méthodes de selection de gaussiennes est d'identifier les gaussiennes pertinentes pour limiter le nombre de vraisemblances à calculer.

3. Algorithmes de selection de gaussiennes

Différentes approches ont été développées pour la selection des gaussiennes contribuant de façon significative au calcul de la vraisemblance d'une observation sachant les mixtures des modèles. Ortmanns & Al [7] proposent un partionnement hiérarchique de l'espace de représentation en hypercubes imbriqués. Il obtient une réduction du temps de calcul d'un facteur 4 ; en combinant cette méthode avec un algorithme de quantification vectorielle, il obtient une réduction d'un facteur 8 sur un système de reconnaissance 20000 mots. Duchateau [2] utilise une méthode d'élimination des gaussiennes basée sur une décomposition des distributions multi-dimensionelles en gaussiennes uni-dimensionnelles. Le processus de sélection opère sur ces dernières par élimination des gaussiennes dont une des composantes au moins est jugée non-pertinente. Cette méthode est fondée sur l'hypothèse d'indépendance des coefficients. Evaluée sur une système 20000 mots comportant 20000 gaussiennes, elle permet une réduction du nombre de distributions selectionnées d'un facteur 22 sans dégradation des taux de reconnaissance.

Bocchieri [1] propose une pré-classification des gaussiennes puis une recherche des classes pertinentes par un algorithme de quantification vectorielle. Pour limiter la représentation incomplète des trames situées aux frontières des classes, il utilise un seuil sur la distance des gaussiennes aux centres qui permet un recouvrement des classes améliorant le processus d'association trame-gaussiennes. Cette méthode a fait l'objet d'une évaluation complète dans [3] ; elle réduit le nombre de gaussiennes d'un facteur 5 sur un système grand vocabulaire sans augmentation significative des taux d'erreurs. Nous avons développé une approche fondée sur le même principe de classification de gaussiennes. Cependant, alors que Bocchieri ne selectionne que la meilleure classe après en avoir augmenté artificiellement les variances, nous proposons de selectionner un nombre variable de classes en estimant, par un algorithme rapide, l'erreur d'approximation des vraisemblances résultant de l'élimination des gaussiennes non-significatives.

4. Partionnement dynamique des distributions

Le processus de partionnement comporte deux étapes distinctes ; dans un premier temps, les gaussiennes sont regroupées en classes et un modèle monogaussien de chaque classe est estimé. Cette première étape est réalisée à l'issue de l'estimation des modèles. Le deuxième processus consiste à selectionner les classes de FDP modélisant de façon significative le voisinage de l'observation. Cette selection est faite en fixant une vraisemblance relative minimale de la trame sachant les modèles de classes. Enfin, seule les émissions des gaussiennes appartenant aux classes selectionnées sont caculées.

4.1. *Classification non-supervisée de gaussiennes*

Nous utilisons l'algorithme des k-means classique ; la distance entre les gaussiennes $N_i(\mu_i, \nu_i)$ et $N_j(\mu_j, \nu_j)$ est une mesure de divergence de Kullback-Leibler symétrisée :

$$D(N_i, N_j) = \sum_{k=1}^{l} \frac{\nu_i^k}{\nu_j^k} + \frac{\nu_j^k}{\nu_i^k} + \frac{(\mu_i^k - \mu_j^k)^2}{\nu_i^k + \nu_j^k}$$

où l est la dimension des vecteurs acoustiques. Un point essentiel pour la convergence de l'algorithme de classification est la cohérence entre la distance choisie et l'algorithme de fusion de gaussiennes : le centre de la classe doit être l'élément minimisant l'inertie de la classe, au sens de la distance utilisée. Les centres des classes résultent de la fusion des p individus de la classe. La FDP obtenue est en fait la gaussienne $N^c(\mu^c, \nu^c)$ modélisant l'ensemble des observations émises par chaque distribution. En notant $N_i^c(\mu_i^c, \nu_i^c), 1 \leq i \leq p$ les FDP de la classe c, on a :

$$\mu^c = \frac{1}{p} \sum_{i=1}^{p} \mu_i^c$$
$$\nu^c = \frac{1}{p} \sum_{i=1}^{p} \left(\mu_i^c \mu_i^{c^T} + \nu_i^c \right) - \mu^c . \mu^{c^T}$$

Les centres de classes constituent les modèles monogaussiens des classes qui vont être utilisés dans la phase de décodage proprement dite.

4.2. *Sélection dynamique des classes de gaussiennes*

Pour chaque observation X, toutes les vraisemblances $P(N^c|X)$ sont calculées. Les trames situées aux frontières des classes étant très mal modélisées par un seul groupe de gaussiennes, nous cherchons toutes le classes significatives pour une observation donnée. Pour cela, les m classes C_i sont ordonnées suivant la relation :

$$P(N^{c_j}|X_i) \geq P(N^{c_{j+1}}, |X_i) \ , \ j \in \{1, m\}$$

Nous retenons q classes parmi les m possibles en fonction d'un rapport de vraisemblance résiduel S fixé à priori. q est le nombre minimal de classes pour lequel la contrainte suivante est respectée :

$$\frac{\sum_{j=1}^{q} P(N^{c_j}|X_i)}{\sum_{j=1}^{m} P(N^{c_j}|X_i)} \geq 1 - S$$

Le calcul des $P(N_j|X_i)$ est ensuite limité aux gaussiennes des classes retenues. L'algorithme de décodage exploitant le partitionnement dynamique comporte donc les 4 étapes suivantes :

- (1) classification non-supervisée de toutes les gaussiennes des MMC en m classes

- (2) estimation des modèles mono-gaussiens des classes de gaussiennes
- (3) pour chaque trame X, calcul et ordonnancement des vraisemblances de X pour chaque modèle de classe
- (4) calcul des émissions de chaque état limité à l'ensemble des gaussiennes des classes seletionnées

5. Système de référence

Nous avons évalué l'apport du partionnement dynamique sur un DAP markovien classique en utilisant le corpus de parole continue TIMIT. Cette base de données contient les enregistrements de 630 locuteurs américains prononçant 10 phrases. 8 phrases des 462 locuteurs sont utilisées pour l'apprentissage, soit 3696 enregistrements. Le corpus de test est constitué de 8 enregistrements de 168 locuteurs, différents des locuteurs de l'apprentissage. Un corpus de test réduit (C-Test) est composé des enregistrements de 24 des 168 locuteurs du corpus de test.

Notre système de référence contient 48 modèles, qui sont ensuite regroupés en 39 classes pour l'évaluation des taux d'erreurs. Ces choix sont classiques et correspondent à ceux proposés dans [5].

Ces modèles contiennent 3 états émetteurs ; seul le passage par l'état central est imposé. Les vecteurs acoustiques sont composés des 12 premiers coefficients cepstraux, de l'énergie, et des différentielles du premier et du second ordre de ces 13 coefficients. Chaque modèle contient au plus 8 gaussiennes ; nous utilisons des matrices de covariances diagonales.

La base TIMIT contient un étiquetage phonétique des phrases de référence. L'alignement initial en état est fait ici par un algorithme d'alignement dynamique qui permet d'obtenir la segmentation d'une séquence d'observations minimisant la somme des variances des segments ; les alignements suivants sont obtenus par un décodage contraint par la séquence phonétique de référence.

L'estimation des mixtures de gaussiennes est effectuée par l'algorithme EM, la mixture initiale étant issue d'une classification des trames en 8 classes par un K-Moyennes. La mixture de 8 gaussiennes correspondante est ensuite re-estimée par EM jusqu'a stabilisation de la vraisemblance moyenne des observations sachant le modèle ; enfin, les gaussiennes qui émettent moins de 10 trames sont supprimées. La suppression d'un élément de la mixture est systématiquement suivie d'une ré-estimation de l'ensemble des gaussiennes. Les probabilités de transitions bigrammes entre modèles sont estimées sur le corpus d'apprentissage avec l'outil du CMU *Statistical Langage Modeling Toolkit.*

La table 1 donne les résultats obtenus sur les corpus de test réduit (C-Test de 192 phrases) et complet (Test de 1344 phrases).

Ces résultats sont légérement meilleurs que ceux obtenus par les systèmes Markoviens de complexité similaire qui ont été développés dans différents laboratoires ([10],[5]), et nettement inférieurs à ceux des modèles plus complexes intégrant notament des informations contextuelles ([4],[10], etc.).

Table 1: Résultats du système de référence sur TIMIT (Test et C-Test)

Corpus	Corr	Sub	Del	Ins	Err
Test	68.1%	21.4%	10.5%	3.6%	35.5%
C-Test	67.5%	21.0%	11.5%	3.2%	35.7%

6. Expériences

Nous avons évalué deux stratégies ; la première consiste à réaliser l'apprentissage de façon classique, sans utiliser le partionnement dynamique pour la segmentation à partir du modèle estimé lors de la passe précédente ; le décodage est ensuite réalisé en calculant les émissions avec partionnement de l'ensemble des gaussiennes. Les résultats sont évalués en terme de durée du calcul des émissions et des taux d'erreurs du DAP. La deuxième stratégie consiste à utiliser le partionnement dynamique à chaque itération de l'algorithme d'apprentissage pour la segmentation du corpus à partir du modèle déjà estimé.

6.1. Décodage

Pour cette première éxpérience nous avons utilisé un modèle appris en trois itérations sans partitionement des gaussiennes. Nous avons ensuite réalisé plusieurs décodages du corpus de test réduit de TIMIT avec partionnement, en faisant varier le seuil du rapport de vraisemblance et le nombre de classes initiales. Nous avons observé l'évolution des taux d'erreurs ainsi que celle de la vitesse du décodage. Cette dernière a été mesurée à posteriori ; en effet, si la méthode proposée doit réduire le nombre d'émissions calculées, elles induit aussi un surcoût lié au calcul des vraisemblances sachant les modèles de classes et au tri des classes suivant l'ordre des vraisemblances calculées. Ce dernier point étant difficilement estimable, nous avons mesuré empiriquement la durée totale du calcul des émissions avec et sans partitionnement. Cette mesure globale tient compte du cot suplémentaire induit par la selection dynamique des classes.

Table 2: Durée totale de décodage et taux d'erreurs en fonction du rapport de vraisemblance résiduelle S pour 8 classes (C-Test)

S	10^{-1}	10^{-2}	10^{-4}	10^{-6}	10^{-8}	0.0
Err. %	39.3	37.0	35.8	35.6	35.7	35.7
Tps.(s)	132	177	283	371	438	890

Table 3: Durée totale de décodage et taux d'erreurs en fonction du rapport de vraisemblance résiduelle pour 64 classes (C-Test)

S	10^{-1}	10^{-2}	10^{-4}	10^{-6}	10^{-8}	0.0
Err.%	51.7	40.1	36.0	35.8	35.7	35.7
Tps.(s)	77	105	169	261	338	890

Les résultats obtenus montrent une réduction de la durée du décodage d'un facteur variant de 2.0 à

Table 4: Durée totale de décodage et taux d'erreurs en fonction du rapport de vraisemblance résiduelle pour 128 classes (C-Test)

S	10^{-1}	10^{-2}	10^{-4}	10^{-6}	10^{-8}	0.0
Err.%	65.0	45.6	36.3	36.0	35.8	35.7
Tps.(s)	104	125	190	261	268	890

11.5 pour une dégradation des taux de reconnaissance comprise entre 0% et 15.4% en valeur absolue (pour 8 classes et $S = 0.1$). Avec une vraisemblance résiduelle de 10^{-6} et avec 64 classes de gaussiennes, on obtient un triplement de la vitesse de décodage sans dégradation des résultats. Globalement, on peut obtenir des résultats équivalents au score de référence en multipliant la vitesse de décodage par 3.3 ; en acceptant un taux de reconnaissance de 36.3% (soit une perte de 0.6% en valeur absolue), on multiplie la vitesse de décodage par 7.1 avec une partition initiale de 128 classes et une vraisemblance résiduelle de 10^{-4}.

6.2. Partionnement dynamique pour l'apprentissage

Nous avons ensuite évalué l'apport de la méthode proposée pour l'estimation des modèles acoustiques. Le décodage contraint est réalisé en utilisant le calcul des émissions basé sur le partionnement des classes de gaussiennes. Nous avons utilisé, pour ce test, une classification initiale de 64 classes et un rapport de vraisemblance résiduelle de 10^{-6}. Nos résultats sur les corpus de test réduit et complets montrent des taux d'erreurs identiques ; le calcul des émissions est, quant à lui, accéléré dans les même proportions (de l'ordre de 340%) que celles observées lors de l'expérience précédente dans une configuration identique.

Table 5: Résultats du système avec partitionnement pour l'estimation (Test et C-Test)

Corpus	Corr	Sub	Del	Ins	Err
Test	67.7%	21.2%	11.1%	3.4%	35.7%
C-Test	68.0%	21.4%	10.6%	3.5%	35.5%

7. Conclusion

L'algorithme de partionnement dynamique des gaussiennes décrit dans cet article permet de réduire considérablement le coût du décodage et de l'apprentissage des modèles acoustiques basés sur des MMC et des mixtures de gaussiennes. L'accélération du décodage à taux de reconnaissance constant est équivallente à celle obtenue par quantification vectorielle, et inférieure à celle publiée dans [2] avec un système de référence de 20000 gausiennes. Cependant, les conditions d'évaluation de ces méthodes sont très différentes des nôtres: d'une part, les résultats sont ici exprimés en terme de temps de calcul effectif, incluant le processus de sélection lui-même; d'autre part, l'exploitation de l'information linguistique réduit la sensibilité du décodeur aux altérations des scores acoustiques. Inversement, l'utilisation d'heuristiques limitant l'exploration de l'ensemble des hyptohèses peut limiter le gain obtenu par la selection des gaussiennes ; enfin, le nombre total de gaussiennes a une influence directe sur le coût total du calcul des émissions ainsi que sur le gain potentiellement obtenu par sélection : en augmentant le nombre de modèles et la complexité des mixtures, on augmente à la fois la précision et la couverture de l'espace par le modèle. Le rapport entre le nombre moyen de gaussiennes significatives n'est sans doute pas proportionnel au nombre total de gaussiennes du modèle, comme le montrent [3] et [2].

Bibliographie

[1] E. Bocchieri. Vector quantization for the efficient computation of continuous density liklihood. In IEEE, editor, *Proc ICASSP'93*, volume 2, pages 692–696, Speech Research Dept.,ATT Lab., Murray Hill, 1993. IEEE.

[2] J. Duchateau. *HMM based acoustic modelling in large vocabulary speech recognition.* PhD thesis, Katholieke Universiteit Leuven, 1998.

[3] K.M. Knill, M.J. Gales, and S.J. Young. Use of gaussian selection in large vocabulary continuous speech recognition using hmms. In *Proc. ICSLP'96*, volume 1, pages 470–474, Philadelphia, PA, USA, 1996. Cambridge University.

[4] L.F. Lamel and J.L. Gauvain. High performance speaker-independant phone recognition using cdhmm. In ESCA, editor, *Proc. Eurospeech*, pages 121–124. LIMSI, 1993.

[5] K.F. Lee and H.W. Hon. Speaker independant phone recognition using hidden markov models. *IEEE Trans. Acoust., Speech and Signal Proc.*, 11:1641–1648, 1989.

[6] F. Lefèvre. *Estimation de probabilité nonparamétrique pour la reconnaissance Markovienne de la parolee.* PhD thesis, Unversité Pierre et Marie Curie, Paris VI, 2000.

[7] S. Ortmanns, T. Firzlaff, and H. Ney. Fast likelihood computation for continuous mixture densities in large vocabulary speech recognition. In *Proc. EUROSPEECH'99.* ESCA, 1999.

[8] M.K. Ravishankar. *Efficient Algorithms for Speech Recognition.* PhD thesis, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, May 1996.

[9] T. Robinson, M. Hochberg, and S. Renals. Ipa: improved phone modelling with recurrent neural networks. *IEEE Trans. on Neural Networks*, 1994.

[10] S.H. Young and P.C. Woodland. State clustering in hidden markov models-based continuous speech recognition. *Computer Speech and Langage*, 8:369–383, 1994.

Utilisation combinée d'indices acoustiques et articulatoires pour la reconnaissance automatique de la parole

Nicolas Petit et Alain Soquet

Laboratoire de Phonologie
Université Libre de Bruxelles
Tél.: ++32 2 650 20 18 - Fax: ++32 2 650 20 07
E-mail: asoquet@ulb.ac.be - http://www.ulb.ac.be/philo/phonolab

ABSTRACT

In this paper, we discuss the role of different complementary cues at three different levels: acoustic, aerodynamic and articulatory on a speech recognition task. The recognition system is a standard speaker dependant HMM recognizer. The corpus is based on CVCV sequences uttered by one speaker. It is shown that aerodynamic and articulatory cues used in combination with Mel frequency cepstral coefficients lead to a substantial increase of the system recognition score.

1. INTRODUCTION

Une alternative pour tenter d'améliorer les performances des systèmes de reconnaissance automatique de la parole est de chercher à compléter les indices acoustiques traditionnels par des informations de type phonologique [Lah99], articulatoire [Zlo93] ou audiovisuelle.

Le but de cette étude est de tester différents indices acoustiques, aérodynamiques et articulatoires dans un système de reconnaissance de mots isolés, et de tenter de déterminer leurs intérêts respectifs.

2. MATÉRIEL

Mesures

Les données ont été acquises sur une station Physiologia [Tes90]. Nous avons mesuré simultanément le signal de parole, la pression intra-orale, les contacts langue-palais et la position de trois capteurs collés sur le visage du locuteur.

Le signal de parole a été échantillonné à 16 kHz dans des conditions équivalentes à une salle d'ordinateurs. Les données électropalatographiques ont été obtenues par le système EPG de l'Université de Reading [Har84]. La pression intra-orale a été mesurée au moyen d'un cathéter de 2 mm de diamètre passé par le consuit nasal jusque dans l'oropharynx (voir figure 1). Ce cathéter était connecté à un capteur de pression de la station *Physiologia*. Le système d'électromagnéto-métrie est le *Movetrack* [Bra85]. Trois capteurs ont été utilisés pour recueillir des informations sur le mouvement des lèvres supérieure et inférieure, ainsi que le menton (voir figure 1).

Corpus

Le corpus est constitué de séquences consonne-voyelle-consonne-voyelle ($C_1V_1C_2V_2$), avec C_1 et C_2 des consonnes parmi [p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, j, ʀ], et V_1 et V_2 des voyelles parmi [i, e, a, o, u, y]. Le nombre total de combinaisons possibles est de 9216. L'objectif était de faire l'acquisition d'un corpus d'une taille raisonnable, tout en couvrant un maximum du nombre total de combinaisons. Nous avons dès lors défini 4 ensembles de 384 séquences de manière à ce que les couples C_1V_1, V_1C_2, et C_2V_2 soient distribués de manière similaire à l'intérieur de chaque ensemble et entre les différents ensembles.

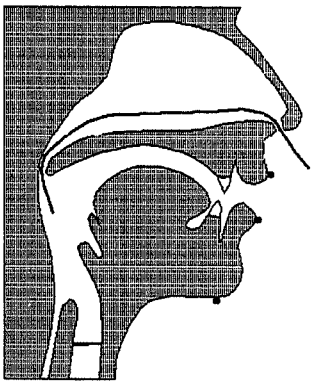


Figure 1: Positionnement du cathéter pour la mesure de la pression intra-orale, et des capteurs pour le *Movetrack*.

3. MÉTHODE

Indices

Les indices ont été calculés tous les 10 ms. Ils ont été calculés à l'instant correspondant au centre de la fenêtre utilisée pour l'analyse acoustique.

Acoustique (m): 12 coefficients cepstraux Mel ont été calculés sur une fenêtre de 25.6 ms du signal de parole [Dav88].

Electropalatgraphique (e): Trois indices ont été extraits des données EPG [Rec93]. Deux indices sont définis sur la dimension antério-postérieure (les coefficients d'antériorité **ca** et de postériorité **cp**), et un sur la dimension latéral-central (le coefficient de centralité **cc**).

Aérodynamique (p): La mesure de la pression intra-orale a été sous-échantillonnée à 100 Hz et utilisée directement comme indice.

Electromagnétométrique: Deux ensembles d'indices basés sur les données EMA ont été utilisés. Le premier ensemble (**a**) est constitué de six dimensions reprenant les coordonnées x et y des trois pastilles (lèvre supérieure **sl**, lèvre inférieure **il**, et menton **j**). Le second (**n**) est constitué de trois grandeurs: l'ouverture aux lèvres **lo** (distance entre les capteurs des lèvres inférieures et supérieures), la protrusion labiale **lp** (moyenne des abscisses des capteurs des lèvres) et l'ouverture de la mâchoire **ja** (angle entre l'horizontale et une droite passant par le capteur du menton et une estimation du point d'articulation de la mâchoire).

Dérivée (d...): Les dérivées temporelles des indices sont calculées grâce à l'expression suivante [You98]:

$$\Delta I_t = \frac{\sum_{\tau=1}^{D} \tau (I_{t+\tau} - I_{t-\tau})}{2\sum_{\tau=1}^{D} \tau^2} \text{ avec D égal à 2.}$$

Vitesse (v...): La norme de la vitesse des vecteurs formés par les indices d'une même classe a également été calculée. Elle correspond simplement à la racine carrée de la somme des carrés des dérivées de ces indices. Ainsi, l'indice **vm** nous renseigne, par son amplitude, sur le taux de variation global des indices de la classe **m**.

Chaînes de Markov

Chaque phonème est modélisé par un modèle de Markov constitué de trois états identiques. Pour chaque état, la probabilité d'émission est générée par une gaussienne multidimensionnelle avec une matrice de covariance diagonale. Une distinction est faite entre les phonèmes situés en début et en fin de séquence, ainsi qu'entre le silence de début et de fin.

Nous avons utilisé l'algorithme de Viterbi [Jua91] pour l'estimation des paramètres des modèles et pour la phase de reconnaissance. Notons qu'aucun des ensembles d'apprentissage n'a été segmenté au préalable.

Sélection d'indices

Phase d'apprentissage: Il est envisageable de procéder à une évaluation de l'efficacité des indices utilisés. On peut considérer trois types d'évaluations :

- Une évaluation globale de l'indice quel que soit le phonème concerné.
- Une évaluation par phonème : dans quelle mesure cet indice distingue-t-il bien tel phonème de tous les autres phonèmes ?
- Une évaluation par couple de phonèmes : dans ce cas, on cherche à évaluer la capacité d'un indice à distinguer deux phonèmes donnés.

Nous proposons ici une réalisation de cette dernière méthode. Afin d'établir une évaluation suffisamment fine, il est nécessaire de confronter l'indice étudié à une multitude de cas. La méthode retenue procède donc de la manière suivante :

- Pour chaque enregistrement du set d'apprentissage, on effectue une segmentation réalisée avec les indices acoustiques (segmentation acoustique).
- Ensuite, pour chacun des phonèmes p_n de cet enregistrement et pour chaque indice I_k, on calcule la probabilité $P(p_n, I_k)$ d'émission du segment correspondant au phonème p_n par l'indice I_k (produit des probabilités gaussiennes sur les vecteurs du segment).
- Enfin, on effectue le même calcul mais avec tous les autres phonèmes p_m de l'alphabet sur le segment correspondant à p_n. Si $P(p_m, I_k) < P(p_n, I_k)$, les scores $S_{m,n,k}$, $S_{n,m,k}$, $S_{n,n,k}$ et $S_{m,m,k}$ sont incrémentés.

En définitive, on crée une base de donnée où, pour chaque couple de phonèmes $\{p_m, p_n\}$ ($m \neq n$), il existe un classement des indices selon leur capacité à différencier p_m de p_n et pour les couples $\{p_m, p_m\}$, un classement des indices distinguant le mieux p_m de tous les autres phonèmes. Pour chaque couple de voyelles, les cotations ont été réalisées à partir de 384 segments et pour les consonnes, à partir de 144 segments.

Le tableau qui suit nous montre pour quelques couples le classement d'indices obtenu.

Table 1: Classement d'indices pour quelques couples.

	{a,u}	%	{b,d}	%	{f,v}	%	{m,n}	%
1	m1	99.48	ca	97.92	p	93.75	ca	97.22
2	ily	99.22	cc	97.22	d m3	82.64	cc	96.53
3	d jx	99.22	d cc	97.22	m4	81.94	ve	95.14
4	lo	98.96	sly	94.44	m5	81.94	d ca	94.44
5	ilx	98.44	v e	94.44	d m1	81.94	d cc	93.75
6	jx	97.92	d ca	89.58	m0	81.25	lo	89.58
...	...	...	...	...	...	...	...	...
56	cc	49.74	m0	50	d ja	50	p	54.86
57	ca	49.74	d m2	49.31	sly	47.92	m10	54.17
58	d ca	49.48	d m1	47.22	v il	47.22	d m9	53.47

On note par exemple l'importance de certains indices dérivés de l'EPG (**ca**, **cc**, **ve**, **d ca** et **d cc**) pour le couple {m, n} ou l'intérêt des indices **ily**, **d jx**, **lo**, **ilx** et **jx**, pour le couple {a, u}.

Une table similaire aurait pu être créée à partir des distributions gaussiennes de l'alphabet. Cependant ici, l'évaluation tient compte des distributions réelles des indices.

Phase de reconnaissance:

Considérons une tâche de reconnaissance effectuée sur le mot « tami ».

Dans un premier temps, une probabilité d'émission P_i est calculée par l'algorithme de Viterbi pour chaque mot du

vocabulaire. À partir d'ici, on peut se contenter de choisir la transcription ayant obtenu la meilleure probabilité, ce qui correspond à une reconnaissance classique. Cependant, nous allons sélectionner les N meilleures transcriptions et effectuer une étude plus poussée de celles-ci à l'aide de la table de classement des indices précédemment obtenus.

Afin d'exploiter notre base de données, nous procédons ensuite en considérant un couple de transcriptions prises parmi ces N meilleurs : par exemple *dami* et *tani*. Les mots ayant même longueur dans le présent corpus, on identifie donc immédiatement les couples à considérer : {d,t}, {a,a}, {m,n} et {i,i}. Pour départager ces deux mots, on va sélectionner l'ensemble formé par les 3 meilleurs indices de chaque couple, ce qui correspond ici à (au plus) 12 indices. L'algorithme de Viterbi nous fourni alors les probabilités $P_{s,1}$ et $P_{s,2}$ respectivement pour *dami* et *tani*. Ces mots ayant obtenu précédemment les probabilités P_1 et P_2, celui des deux qui obtient le produit $P_i * P_{s,i}$ le plus élevé remporte le « match ».

On retient enfin celui, parmi ces N candidats, qui remporte le plus de matchs.

Comme on le voit, il faut organiser N(N-1)/2 matchs, soit N(N-1) recours à l'algorithme de Viterbi, ce qui correspond, pour N = 6, à moins de 2% de la taille du vocabulaire. Le surplus d'opérations se révèle donc presque négligeable.

Les comparaisons 2 à 2 permettent de limiter le nombre d'indices sélectionnés lors de chaque mise à l'épreuve d'une transcription. Puisque l'on ne peut comparer des probabilités obtenues avec des indices différents, c'est le nombre de matchs remportés qui devient critère de sélection global.

Un point essentiel dans cette discussion est le choix des indices responsables de la segmentation. Il est en effet important de garder un même type de segmentation. Lors d'un match, le nombre d'indices sélectionnés peut être faible lorsque les mêmes indices sont retenus par plusieurs couples. De plus, en fonction des indices choisis (acoustiques ou articulatoires), l'algorithme de Viterbi proposera des parcours (et donc des segmentations) fort variables. Il est donc préférable de calculer les probabilités de ces indices sur base de la segmentation acoustique. Ainsi, la pression intra-orale moyenne d'un *t* n'est mesurée que lorsqu'il est audible même si la pression croît avant le *burst* final.

Avec des mots de longueurs variables, la méthode reste applicable, il est cependant nécessaire d'établir, grâce à la segmentation de chaque mot, quels sont les couples de phonèmes formés pour chaque vecteur de l'enregistrement.

Expériences

Pour chaque expérience, une technique de jack-knife a été utilisée. Le corpus étant composé de quatre ensembles bien équilibrés, chaque ensemble a été utilisé comme ensemble de test et les trois ensembles restants comme ensembles d'apprentissage.

Le système est évalué sur une tâche de reconnaissance de mots parmi les 1536 mots du corpus.

4. RÉSULTATS

La figure 1 (barres blanches) montre la distribution de la place qu'occupe la bonne transcription lorsque la reconnaissance est effectuée avec les seuls indices de la classe **m**. Le score du rang 1 correspond donc au score de reconnaissance obtenu avec ce type d'indices.

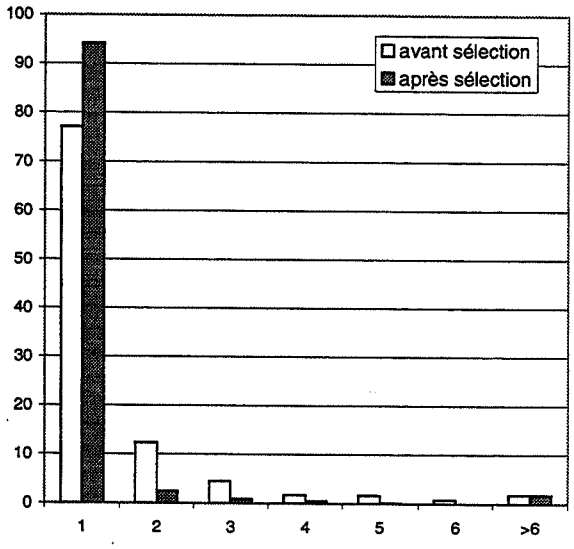


Figure 1: Performances de reconnaissance de mots en fonction du rang, avant et après sélection des indices.

En choisissant N = 6, on a l'assurance que, dans 98.2 % des cas, la transcription correcte participera à la compétition.

Table 2: Classement des indices en fonction de l'efficacité globale.

1	m1	88.7	15	lo	75.5	29	lp	69.0	41	d jy	65.5
2	m0	86.4	16	d cc	75.5	30	d m7	68.5	42	v il	65.4
3	m2	84.2	17	cp	74.5	31	d m2	68.3	43	v sl	65.0
4	m3	82.2	18	m11	74.0	32	ilx	67.9	44	d cp	64.9
5	p	80.9	19	d m1	73.6	33	d m9	67.7	45	v j	64.3
6	cc	80.4	20	d m0	73.2	34	d m5	67.6	46	d m11	64.1
7	m4	79.6	21	v e	72.8	35	d sly	67.4	47	d lp	64.1
8	m6	79.5	22	ily	72.8	36	d m8	67.0	48	d ja	63.9
9	m10	78.5	23	jx	72.2	37	d m4	66.6	49	jy	63.9
10	m7	78.1	24	m9	72.1	38	d jx	66.6	50	d slx	63.6
11	m5	77.0	25	d m3	70.2	39	d ilx	66.5	51	v n	63.3
12	v m	76.4	26	sly	70.0	40	ja	66.4	52	d m6	63.0
13	d p	76.3	27	d ca	69.7	41	d lo	66.0	53	slx	62.5
14	m8	75.9	28	ca	69.2	42	d ily	65.8	54	d m10	62.5

Les résultats avec sélection d'indice et compétition entre les 6 meilleures transcriptions sont présentés à la figure 1 (en gris).

On peut maintenant obtenir une évaluation grossière de l'efficacité globale de chaque indice à partir de la table des meilleurs indices en sommant pour chaque indice I_n les $S_{n,n,k}$ sur les phonèmes p_k (voir table 2). Ce classement reste approximatif car une somme est effectuée faisant intervenir des couples dont certains ne posent jamais problème et d'autres comme {d,b} qui donnent lieu à de nombreuses erreurs de reconnaissance.

Une reconnaissance classique a ensuite été réalisée avec un nombre N croissant des meilleurs indices, le graphique qui suit résume les scores obtenus. Notons que, contrairement aux résultats précédents, la segmentation n'est plus totalement acoustique.

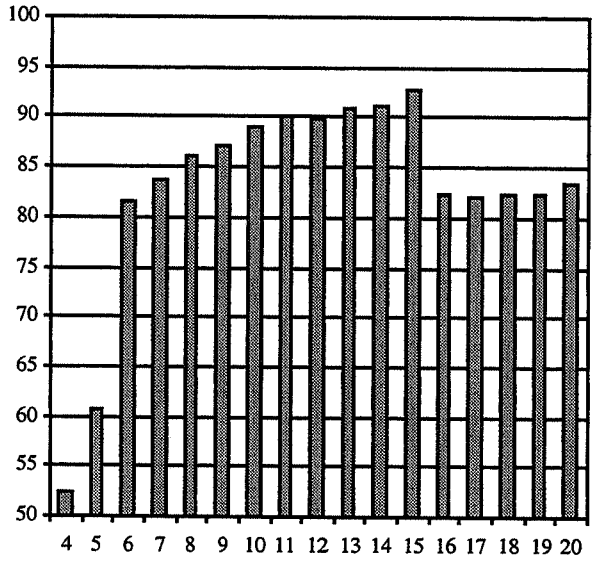


Figure 2: Score de reconnaissance de mots [%] en fonction du nombre d'indices.

La brusque dégradation observée pour N=16 peut s'expliquer par le fait que l'indice **d cc** (16ème du classement), de par sa nature de dérivée, est mauvais pour segmenter, même s'il se révèle bon avec une segmentation acoustique. À mesure que le système inclut un nombre croissant d'indices non-acoustiques, on bascule dans un autre type de segmentation.

Le tableau suivant résume les scores obtenus lors de nos expériences ainsi que le nombre N d'indices utilisés lors de la phase de reconnaissance. Il met en évidence d'une part l'intérêt de l'utilisation d'indices articulatoires et d'autre part le gain appréciable que procure la sélection d'indice globale et celle par couple.

Table 3: Performances de reconnaissance.

	m	**me**	**mea**	**mepn vm**	**sel. globale**	**sel. par couple**
N	12	15	21	20	15	3 à 12
%	77.1	80.1	85.1	89.3	92.7	94.1

5. CONCLUSIONS

Dans ce travail, nous avons testé différents indices acoustiques, aérodynamiques et articulatoires dans un système de reconnaissance de mots isolés.

La procédure de sélection d'indice augmente les performances du système et fournit un ensemble d'informations intéressantes sur le choix d'indices à effectuer pour discriminer au mieux un couple de phonème.

BIBLIOGRAPHIE

[Bra85] Branderud, P. 1985. Movetrack – a movement tracking system. In Proceedings of the French Swedish Symposium on Speech, GALF, Grenoble, France, 113-122.

[Cho90] Chow, Y. et Schwartz, R. 1990. the N-Best algorithm: an effecient procedure for finding top N sentence hypothesis. Proc. ICASSP-90. pp.81-84.

[Dav80] Davis, S. et Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE ASSP-28, 4: 357-366.

[Har84] Hardcastle, W. 1984. New methods of profiling lingual-palatal contact patterns with electropalatography. Speech Research Laboratory Work in Progress, University of Reading, United Kingdom. 4: 1-40.

[Jua91] Juang, B.H. et L.R. Rabiner.1991. Hidden Markov models for speech recognition. Technometrics, Vol.33, 3, 251-272.

[Lah99] Lahiri, A. 1999. Speech recognition with phonological features. Proc. ICPhS-99. San Francisco. pp. 715-718.

[Rec93] Recasens, D., et al. 1993. An electropalatographic study of stop consanant clusters. Speech Communication, 12: 335-355.

[Tes90] Teston, B. et B. Galindo. 1990. Une station de travail d'analyse de la production de la parole. 18èmes Journées d'Etude sur la Parole, Montreal, 1990, 180-184.

[You98] Young, S. 1998. Acoustic Modelling for Large Vocabulary Continuous Speech Recognition, Dans *Computational Models of Speech Pattern Precessing*, dans NATO ASI *series F*, pp 18-39, Springer-Verlag, Berlin.

[Zlo93] Zlokarnik, I. 1993. Experiments with an articulatory speech recognizer. Proceedings of Eurospeech, Berlin, Germany, 2215-2218.

REMERCIEMENTS

Cette recherche a été subventionnée par la convention ARC "Dynamique des système phonologiques" 98-02, n°226.

Sélection dynamique de modèles de langage dans une application de dialogue

Y. Estève, F. Béchet, R. De Mori

LIA - CERI
Université d'Avignon, BP 1228
84911, Avignon Cedex 9, France
Tél.: ++33 (0)490 84 35 00 - Fax: ++33 (0)490 84 35 01
http://www.lia.univ-avignon.fr

Abstract

We propose, in this paper, a method which allows us to build statistical language models dedicated to specific dialog situations. We will describe the architecture of a speech recognition system using several language models. The first step, in the system, consists in producing a word-lattice from a given sentence uttered by a speaker. A general language model calculates a sentence-hypothesis. Then, in a second step, the system choose a specialised language model according to the word-lattice and the previous hypothesis. Another decoding process is performed using this specialised language model in order to produce a new sentence-hypothesis. Finally, a decision-module processes these two hypothesis in order to choose one of them or to reject both of them.

1. Introduction

Les serveurs vocaux interactifs sont une des principales applications des systèmes de Traitement Automatique de la Parole, que ce soit en reconnaissance ou en synthèse. Ces serveurs, généralement limités à une tâche particulière représentant un domaine sémantique bien défini, utilisent des lexiques de taille réduite. Cela permet d'assurer des performances suffisantes au module de reconnaissance de parole pour rendre efficace l'utilisation de tels serveurs.

Les modèles de langage utilisés habituellement dans ces modules de reconnaissance sont des modèles statistiques de type n-grams. Ils sont appris sur des corpus d'entraînement représentatifs de la tâche visée. La faible taille du vocabulaire permet l'utilisation de corpus de taille relativement petite pour la phase d'apprentissage, tout en obtenant de bons résultats en terme de perplexité. Des variantes des modèles n-grams classiques ont été proposées dans [5] et [2] afin de regrouper les unités de reconnaissance selon des classes adaptées aux spécificités de la reconnaissance de la parole en situation de dialogue.

Cependant, un examen attentif de retranscriptions de dialogue entre un utilisateur et un serveur vocal permet de segmenter les interventions de l'utilisateur en plusieurs catégories représentatives de la situation du dialogue à un moment donné. Notamment, la variabilité des phrases prononcées par un utilisateur face à des questions directes du système est très faible. Pour modéliser ces "phrases-clefs", il a été proposé de mettre au point des modèles de langage ayant une portée plus grande que les bigrammes ou trigrammes utilisés habituellement ([1]). L'utilisation d'un modèle robuste, indépendant du thème et dépendant du style de dialogue a été présenté dans [7]. En reprenant l'idée de modèles de langage multiples, nous pensons qu'il est intéressant de disposer, en plus d'un modèle de langage général, de modèles de langage spécialisés dans certaines situations de dialogue. Ces modèles spécialisés peuvent avoir deux types d'utilisation : d'une part mieux capter les régularités des interventions d'un utilisateur dans certaines phases du dialogue ; d'autre part permettre, dès la phase de décodage, de rajouter à la phrase reconnue par le module de reconnaissance des informations concernant le type de phrase qui vient d'être prononcée (confirmation, négation, question générale, etc.). Ces informations peuvent permettre d'augmenter la robustesse du module de gestion du dialogue en cas de reconnaissance défaillante.

2. Contexte de l'étude

Cette étude se base sur un corpus de retranscription de dialogue, le corpus AGS du CNET [4], illustrant un dialogue téléphonique entre un utilisateur et un serveur vocal. Ce serveur vocal a pour objectif de guider l'utilisateur vers d'autres serveurs vocaux spécialisés dans une tâche donnée (serveur météo pour une région, serveurs d'emploi, ...). Les données mises à notre disposition par le CNET sont les suivantes :

- le corpus de retranscription de dialogue AGS contenant environ 9800 phrases pour un vocabulaire de 800 mots ;
- un ensemble de 400 phrases de développement avec les graphes de mots issus du système de reconnaissance du CNET (graphe comportant les scores acoustiques pour chaque mot).

Dans le corpus AGS, certaines phrases ont une fréquence d'occurrence très élevée (les 10 phrases les plus fréquentes représentent 20% des occurrences). Ces phrases correspondent généralement à des réponses de l'utilisateur face à une question directe du système. Une étude manuelle du corpus met facilement en évidence le faible nombre de "type" de phrases différentes composant le corpus.

Selon les critères considérés, il est possible de séparer le corpus en sous-groupes représentatifs d'une situation particulière de dialogue. Par exemple on peut

séparer les phrases en utilisant le domaine de la requête (serveur météo ou d'emploi), ou encore la formulation de la requête (par exemple, les phrases commençant par "je voudrais le numéro du serveur ..."). Il est également facile de détecter les réponses faites par l'utilisateur aux propositions du système ("oui", "non", "annulation", etc.).

Selon les critères de sélection employés, la taille et le type des sous-corpus changent énormément. Nous avons choisi une stratégie basée sur les arbres de décision en utilisant comme critère de sélection la réduction de perplexité entre le corpus du noeud père et les sous-corpus obtenus en chaque noeud. Le type d'arbre utilisé est celui des arbres SCT (Semantic Classification Tree) introduit par Kuhn et de Mori [6].

3. Segmentation du corpus d'apprentissage

Les arbres SCT permettent de construire, de manière automatique, des expressions régulières découpant le corpus d'entraînement en sous-corpus. Ces arbres utilisent une liste d'exemples, un ensemble de questions et un critère pour choisir la meilleure question en chaque noeud. Nous allons présenter rapidement ces trois paramètres.

3.1. *Liste d'exemples*

Les exemples utilisés pour construire l'arbre de décision sont les phrases du corpus AGS. Ces phrases ont été étiquetées syntaxiquement en utilisant un tagger développé au LIA, puis lemmatisées par rapport à un dictionnaire de référence.

3.2. *Ensemble de questions*

Dans les arbres SCT, les questions sont générées à partir d'un lexique et d'un ensemble de 3 symboles : $<, >, +$. Les symboles $<, >$ signifient début et fin de phrase et le symbole $+$ représente une suite de mots non vide. Durant la construction de l'arbre, chaque noeud contient une expression régulière, appelée 'Structure Connue' (SC). Cette structure est initialisée, à la racine, avec la valeur :

$SC(racine) = " < + > "$.

Les questions, pour un noeud donné, sont constituées à partir de la valeur SC et du lexique de départ. Ce lexique contient un ensemble de lemmes susceptibles d'être contenus dans les phrases d'exemple. Chaque symbole $+$ de la SC va être remplacé par chaque lemme l_i du lexique de 4 manières différentes : l_i , $+l_i$, l_i+ , $+l_i+$. Il suffit ensuite de tester, pour chaque phrase du corpus d'apprentissage, si elle satisfait ou non l'expression régulière représentée par la SC du noeud. On obtient ainsi deux sous-corpus pour chaque question.

3.3. *Critères de sélection*

Pour pouvoir sélectionner une question parmi toutes celles susceptibles d'être posées en chaque noeud, nous avons choisi d'utiliser le critère de la réduction de perplexité entre le corpus du noeud père et le sous-corpus des phrases ayant répondues positivement à la question posée. A cet effet, nous avons utilisé un sous-ensemble du corpus AGS comme corpus de développement. Ainsi, pour un noeud et une question donnés, nous allons déterminer 4 sous-corpus :

- A_{oui} : corpus des phrases d'apprentissage satisfaisant l'expression régulière ;
- A_{non} : corpus des phrases d'apprentissage ne satisfaisant pas l'expression régulière ;
- D_{oui} : corpus des phrases du développement satisfaisant l'expression régulière ;
- D_{non} : corpus des phrases du développement ne satisfaisant pas l'expression régulière.

Puis nous calculons trois modèles de langage bigrammes : $M_{A_{oui}}$, $M_{A_{non}}$ et $M_{A_{oui}+A_{non}}$ sur les corpus d'apprentissage : A_{oui} , A_{non} et $A_{oui} + A_{non}$.

Enfin nous estimons les perplexités suivantes (avec la fonction de calcul de perplexité PP) :

$P_{oui} = PP(M_{A_{oui}}, D_{oui})$; $P_{non} = PP(M_{A_{non}}, D_{non})$; $P_{ref} = PP(M_{A_{oui}+A_{non}}, D_{oui} + D_{non})$.

La question qui est choisie est celle qui minimise la quantité $P_{oui} + P_{non}$ tout en satisfaisant la condition suivante : $P_{oui} + P_{non} < 2 * (P_{ref} + \varepsilon)$ (où la valeur ε permet de contrôler la taille de l'arbre).

On arrête de faire grandir l'arbre si la taille des sous-corpus devient trop petite où si on ne trouve pas de question qui permettent de satisfaire la condition précédente.

4. Obtention de modèles de langage spécialisés

À chaque noeud de l'arbre SCT obtenu peut être associé un modèle de langage bigramme calculé à partir de toutes les phrases satisfaisant l'expression régulière contenue dans le noeud. Malheureusement, la taille du corpus général étant faible, la petite taille de certains sous-corpus entraîne une spécialisation trop importante du modèle de langage résultant.

Pour éviter ce phénomène nous avons décidé d'utiliser les techniques d'adaptation de modèles de langage à partir d'un petit corpus présentées dans [3]. Dans ce cas, le modèle de langage général est celui appris sur l'ensemble du corpus d'apprentissage et le corpus d'adaptation est le sous-corpus correspondant aux noeuds de l'arbre.

5. Architecture du système

Il paraît difficile de choisir, au moment du décodage, un modèle de langage particulier parmi tout ceux disponibles. Ainsi, nous avons choisi d'utiliser les résultats d'un premier décodage avec le modèle de langage général (appelé LM_1) afin de choisir un modèle spécialisé (appelé LM_2).

L'architecture du système de reconnaissance de la parole proposée ici est décrite figure 1. C'est un système fonctionnant en deux passes : dans un premier temps,

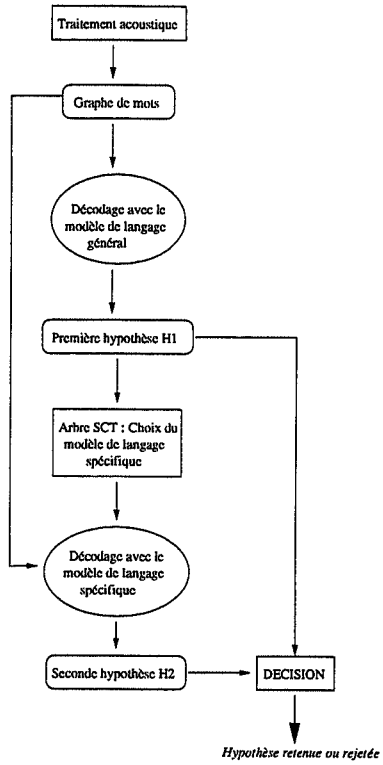


Figure 1: Architecture du système

on calcule le chemin de probabilité maximale dans le graphe d'hypothèses lexicales produit par le système de reconnaissance du CNET avec le modèle LM_1. Le résultat de ce premier décodage, appelé hypothèse H_1, va nous servir de référence dans la seconde passe du système en nous permettant de choisir le modèle LM_2. Ce sous-modèle de langage produira une hypothèse H_2. Nous allons présenter maintenant l'utilisation de ces deux hypothèses.

5.1. *Choix du modèle spécialisé*

Pour choisir le modèle LM_2, nous allons appliquer l'hypothèse H_1 à l'arbre SCT construit lors de l'apprentissage. Ceci permet d'établir à quel sous-corpus spécifique cette hypothèse peut être associée. Les tests effectués montrent que malgré les substitutions, insertions ou suppressions de mots pouvant exister dans la première hypothèse, cette méthode désigne le corpus associé effectivement à la phrase de référence (i.e. la phrase à reconnaître) avec beaucoup de précision.

A l'aide du modèle LM_2, un nouveau décodage est effectué pour produire l'hypothèse H_2. Ces deux hypothèses représentent des informations différentes : si la phrase prononcée est très proche des phrases "types" des situations de dialogue vues dans le corpus d'apprentissage, c'est H_2 qui doit être considérée ; par contre, dans le cas où la phrase représente une demande générale de l'utilisateur, le modèle LM_2 n'a pas un pouvoir de généralisation suffisant pour traiter efficacement le graphe et c'est H_1 qui doit être choisie.

Une phase de décision devient nécessaire. Elle permet ou bien de choisir l'hypothèse à retenir, ou bien de rejeter les deux phrases si certains indices lui indiquent qu'aucune des hypothèses ne peut être correcte.

5.2. *Phase de décision*

Quand les hypothèses H_1 et H_2 désignent des phrases différentes, le système va tout d'abord essayer de s'affranchir des problèmes de réglage du poids des modèles de langage par rapport au modèle acoustique. A cet effet nous effectuons une normalisation des poids des modèles LM_1 et LM_2 afin de réduire la distance (en terme de mots différents, i.e. insertions, substitutions, suppressions) séparant les deux hypothèses.

La phase de décision utilise des critères empiriques portant sur les informations connues de chacune des deux hypothèses de phrase. Ces informations peuvent être leur score acoustique, leur score linguistique, leur score syntaxique, etc.

Ainsi, à partir de ces informations, des règles sont établies qui permettent soit de choisir H_1 ou H_2, soit de rejeter les deux hypothèses: cette dernière décision revient à nier la possibilité de trouver la bonne phrase dans le graphe de mots issu du traitement acoustique.

Voici un exemple d'application de la règle la plus simple (si une hypothèse a un meilleur score acoustique, linguistique, et syntaxique que sa concurrente, alors cette hypothèse est choisie).

H_1: je vous me si ce sera tout (score acoustique:-33751.75, score linguistique: -10.8549, score syntaxique: -30.6634)

H_2: je vous remercie ce sera tout (-33585.3, -10.2012, -20.8851).

Les scores sont sous forme de log10 de probabilités.

Avec cette règle, H_2 sera choisie par le système, car chacun de ses scores est supérieur à son équivalent dans H_1. Bien entendu, quand les hypothèses désignent la même phrase, celle-ci est retenue comme choix du système.

6. Expériences et résultats

Les résultats présentés ici sont obtenus sur les 393 graphes de mots produit par le système du CNET.

Les performances du système initial (système du CNET avec modèle LM_1) sont indiquées dans le tableau 1 en indiquant le taux de graphes incomplets (c'est-à-dire ne contenant pas la phrase de référence), le taux de reconnaissance pour les phrases entières et le taux d'erreurs/mots.

	%	Nb de phrases
Graphes incomplets	41.7	164/393
Phrases correctes	47.3	186/393
erreurs/mots	30.8	

Table 1: Performance du système initial

En rajoutant le modèle spécialisé LM_2 et le module de décision, nous obtenons les résultats présentés dans

le tableau 2.

	%	Nb de phrases
Phrases rejetées	11.95	47/393
Phrases correctes	49.1	193/393
erreurs/mots	30.2	

Table 2: Performance du système global

Les résultats présentés dans le tableau 2 s'interprètent de la manière suivante : tout d'abord le système a rejeté 11.95% des graphes de mots, considérant qu'ils étaient incorrects et qu'il valait mieux faire répéter l'utilisateur plutôt que de diriger le dialogue sur une fausse piste. Il faut noter que tous les graphes rejetés font partie des graphes incomplets mentionnés dans le tableau 1.

Sur les 346 phrases validées par le système, 193 sont correctes, soit 55,7% des phrases validées. Cela représente 49.1% des phrases totales à comparer avec les 47.3% obtenues en utilisant le seul modèle LM_1. On a donc une légère amélioration du taux de reconnaissance par phrase. Par contre, le taux d'erreurs/mots sur les phrases retenues par le module de décision reste équivalent à celui obtenu avec le modèle LM_1.

On peut expliquer cela de la manière suivante : lorsque la phrase à reconnaître correspond aux phrases types modélisées par LM_2, la reconnaissance de la phrase se trouve améliorée ; par contre, dans le cas où les hypothèses acoustiques sont de mauvaise qualité, les règles de décision ont du mal a choisir entre H_1 et H_2. Ainsi, le modèle LM_2 peut être choisi a tort pour décoder une phrase représentant une requête générale, ceci ayant comme effet de dégrader la reconnaissance effectuée par LM_1. En effet, le choix systématique de l'hypothèse H_2 produit un taux d'erreurs/mots de 34.2%, à cause de la faible capacité de généralisation des modèles LM_2.

7. Conclusions et perspectives

L'intéret de la méthode présentée dans cet article se situe à deux niveaux : d'une part lorsque les hypothèses acoustiques sont cohérentes et que les phrases prononcées par l'utilisateur entrent dans un cas typique du dialogue, les modèles spécialisés permettent d'affiner la reconnaissance ; d'autre part le module de décision associe à la transcription de la phrase une étiquette utile au système de gestion du dialogue. En effet cette étiquette représente soit un rejet de la phrase à décoder, soit une situation particulière du dialogue. Cette situation est identifiée en fonction du noeud choisi dans l'arbre SCT pour construire le modèle de langage spécialisé et peut correspondre à une acceptation, un refus, une requête, un commentaire, etc. Le rejet de certaines phrases permet d'éviter que le système de dialogue ne s'égare sur une fausse piste en cas de mauvaise reconnaissance.

Les résultats communiqués ici ne sont que des résultats préliminaires. Le système proposé utilise un grand nombre de paramètres qu'il faut encore optimiser. Par exemple, les règles utilisées lors de la phase de décision ont été fabriquées de manière empirique. L'utilisation d'un arbre de décision sur un corpus de développement pour en construire de plus efficaces est envisagé.

Nous pouvons toutefois constater que les modèles de langage classiques montrent leurs limites lorsque les hypothèses acoustiques sont peu cohérentes. L'utilisation de tels modèles spécialisés dans certaines tâches de dialogue en complément d'un modèle plus général ne permet pas de réparer ces incohérences.

Il est donc intéressant de se pencher sur la notion de mesure de confiance donnée par le système de reconnaissance pour chaque phrase hypothèse ou même pour certaines parties de ces phrases. Ainsi la détection d'îlots de confiance permettrait au système de dialogue d'être plus souple dans son interaction avec l'utilisateur. Il nous faut donc adapter le système de rejet des phrases à un niveau inférieur (par exemple, au niveau des syntagmes : groupes nominaux, verbaux, etc.). En effet, puisque dans des conditions difficiles un grand nombre de phrases reconnues ne sont pas correctes, il est primordial de récupérer dans ces phrases le maximum d'informations fiables afin de guider le système de gestion du dialogue.

Ces travaux sont financés par le centre de recherche de France-Telecom (CNET) sous le contrat 971b427

Bibliographie

[1] Nasr A., Estvève Y., Béchet F., Spriet T., and De Mori R. A language model combining n-grams and stochastic finite state automata. In *Eurospeech*, Budapest, 1999.

[2] Beaujard C. and Jardino M. Un modèle de langage de langage mixte basé sur la similarité des mots dans un système de reconnaissance de la parole. In *JEP*, 1998.

[3] Janiszek D., , De Mori R., Béchet F., Matrouf D., and Mokbel C. New language model adaptation algorithm based on the definition of cardinal distance. In *ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, Kloster-Irsee, Germany, 1999.

[4] Sadek D., Ferrieux A., Cozannet A., Bretier P., Panaget F., and Simonin J. Effective human-computer cooperative spoken dialogue : the ags demonstrator. In *ICSLP*, Philadelphia, 1996.

[5] Damnati G. Modèles de langage et classification automatique pour la reconnaissance de la parole continue dans un contexte de dialogue oral homme-machine. In *Thèse de l'université d'Avignon et des Pays du Vaucluse*, 2000.

[6] Kuhn R. and De Mori R. The application of semantic classification trees to natural language understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(449-460), 1995.

[7] Kawahara T. and Doshita S. Topic independent language model for key-phrase detection and verification. In *ICASSP*, 1999.

Systèmes d'alignement automatique & études de variantes de prononciation

Martine Adda-Decker et Lori Lamel

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay cédex, FRANCE
{lamel,madda}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

This contribution aims at evaluating the use of pronunciation variants across different system configurations and speaking styles in French. The study is limited to the use of variants during speech alignment, given an orthographic transcription and a phonemically represented lexicon, thus focusing on the modeling abilities of the acoustic word models. Parallel and sequential variants are tested in order to measure the spectral and temporal modeling accuracy. To measure the need for variants we have defined the *variant2+* rate which is the percentage of words in the corpus, not aligned with the most common phonemic transcription. Alignment results using different acoustic model sets demonstrate the dependency between acoustic model accuracy and pronunciation variants. A comparison between read and spontaneous speech is presented for French based on alignments from BREF (read) and MASK (spontaneous) data.

1. INTRODUCTION

Les variantes de prononciation peuvent s'expliquer pardifférents facteurs, comme le style de parole, la vitesse d'élocution, des habitudes individuelles ou des accents régionaux... La modélisation de variantes de prononciation pour la reconnaissance automatique de la parole a attiré beaucoup d'intérêt ces dernières années [Spe99] et des exemples pour le français peuvent être trouvés dans [Per98, Mok98], une étude sur l'influence de la vitesse d'élocution dans [Lus98]. L'ajout de variantes de prononciation dans le dictionnaire du système de reconnaissance permet d'accroître les possibilités de modélisation acoustique des mots, et l'effet souhaité est d'arriver à des modèles de mot plus précis et par là à un meilleur décodage. Cependant si les variantes rajoutées ne sont pas pertinentes pour les données acoustiques traitées et/ou par rapport aux faiblesses du décodeur, les performances globales du système peuvent décroître. Combien de fois a-t-on pu observer que les variantes ajoutées n'ont pas permis d'arriver à une amélioration globale: alors que des erreurs sont ponctuellement corrigées, de nouvelles erreurs peuvent s'introduire. Les variantes contribuant à augmenter le taux d'homophones dans le système, elles deviennent des sources d'erreurs potentielles et elles ne sont que peu utilisées dans nos systèmes de reconnaissance [Lam96].

Dans une étude récente [Add99] nous nous sommes intéressés aux variantes possibles lors de simples expériences d'alignement où les dictionnaires de prononciations contiennent un nombre plus ou moins élevé de variantes et où l'alignement n'est guidé que par les modèles acoustiques sans biais du modèle de langage. Nous continuons ici ce travail avec différents corpus en français. L'utilité des variantes est mesurée suivant différents axes: la configuration du système et le style de parole dans les corpus (lu, spontané). On distingue des variantes de prononciation séquentielles et parallèles. Les variantes séquentielles permettent certains phonèmes d'être optionnels ce qui donne une plus grande flexibilité pour la modélisation temporelle des mots. Les variantes parallèles permettent de remplacer un phonème par n'importe quel autre phonème d'un sous-ensemble défini a priori, augmentant ainsi les possibilités de modélisation spectrale. Les variantes observées lors de l'alignement peuvent s'expliquer soit simplement par des faiblesses de modélisation oubien, si la modélisation acoustique est précise, les variantes correspondent à une réalité linguistique et peuvent servir à des études phonétiques.

2. CORPUS DE PAROLE

Deux corpus sont utilisés pour nos expériences. La parole lue provient du corpus BREF [Lam91] qui correspond à la lecture d'articles du journal *LeMonde*. La parole spontanée concerne des demandes d'informations SNCF et a été enregistrée au LIMSI pour le projet ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) [Lam95]. Le contenu de ces corpus est résumé dans la Table 1.

Table 1: Pour chaque corpus sont indiqués le nombre d'énoncés, la durée de parole en nombre d'heures, le nombre total de mots et le nombre de mots distincts.

Corpus	BREF	MASK
style	lu	spontané
#énoncés (distincts)	6.5k	38k
durée parole	120h	35h
#mots(total)	1.1M	260k
#mots(distincts)	25k	2k

La figure 1 montre, pour chaque corpus, la couverture lexicale cumulée en fonction du rang de fréquence des mots. Pour MASK (parole spontané, limitée à un domaine) les 10 mots les plus fréquents représentent 30% des mots du corpus, alors que pour les journaux lus ils couvrent 20%. Avec les 100 mots les plus fréquents 80% des mots du corpus MASK sont couverts, mais seulement 50% pour BREF. Alors que la courbe de BREF est quasi-linéaire sur une échelle logarithmique, la courbe de MASK a une forte pente entre les rangs 10 et 200 et s'aplanit rapidement audelà, ce qui traduit la grande spécificité du corpus.

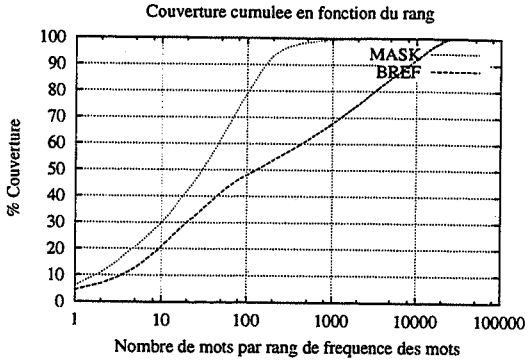


Figure 1: Couverture lexicale en fonction du nombre de mots triés par rang de fréquence du mot pour les corpus spontané (MASK) et lu (BREF).

3. DICTIONNAIRES DE PRONONCIATION

À partir de nos dictionnaires de prononciation standards (référence) nous avons créé des dictionnaires augmentés de variantes parallèles ou séquentielles. Notre but est d'accroître notre compréhension des facultés et limites des modèles acoustiques concernant la modélisation spectrale et temporelle en comparant des résultats d'alignement à travers différents styles de parole.

Dictionnaires de référence Dans la Table 2 on peut voir quelques entrées de notre dictionnaire de prononciation de référence utilisé pour l pprentissage des modèles acoustiques. Ces dictionnaires contiennent typiquement de 10 à 20% de variantes concernant les mots outils fréquents, les nombres ((4) dans la Table 2), les acronymes (5) et les noms propres (6) souvent d'origine étrangère. Un nombre important de variantes concernent le schwa en position finale (3,4) et les liaisons (2,4).

Table 2: Exemples d'entrées lexicales dans le dictionnaire de référence avec variantes parallèles ([]: phonèmes au choix) et séquentielles ({}: phonèmes optionels).

république	repyblik	(1)
les	le{z}	(2)
prendre	prɑ̃dr{ə} prɑ̃d	(3)
dix	dis{ə} di{z}	(4)
DM	d[œ,ə]tSmark dɛm	(5)
Morgan	mɔrgɑ̃ mɔrgɑn	(6)

Dictionnaires à variantes séquentielles Des dictionnaires incluant un très grand nombre de variantes séquentielles ont été dérivées à partir des dictionnaires de référence en rendant une partie des phonèmes optionnels. Ces dictionnaires, appelés *Vopt* (voyelles optionnelles) et *Copt* (consonnes optionnelles) ont pour but de localiser d'éventuels problèmes de modélisation temporelle concernant les modèles acoustiques de mots. Un extrait de ces dictionnaires est montré dans la Table 3. Le phénomène du schwa optionnel en fin de mot est ainsi pris en compte dans les dictionnaire séquentiels. Le dictionnaire *Copt* peut servir à étudier les phénomènes de réduction concernant les clusters de consonnes. Le dictionnaire *Vopt* peut servir à étudier si d'autres voyelles que le schwa sont susceptibles de disparaître et dans quel contexte. De telles omissions, a priori rares, apparaissent assez fréquemment en spontané, accompagnées d'une restructuration syllabique.

Table 3: Exemples d'entrées lexicales dans les dictionnaire *Vopt* et *Copt* montrant une grande flexibilité temporelle. Le schwa final {ə} est optionnel dans tous les dictionnaires.

Vopt	république	r{e}p{y}bl{i}k{ə}
Copt	république	{r}e{p}y{b}{l}i{k}{ə}

Dictionnaires à variantes parallèles Ces dictionnaires ont été générés en définissant des classes de phonèmes et en autorisant un phonème d'une classe donnée à être remplacé par n'importe quel autre membre de cette même classe. Pour chaque classe un dictionnaire spécifique a été créé. La Table 4 montre les classes de phonèmes utilisés dans les travaux décrits ici.

Table 4: Classes de phonèmes pour dictionnaires parallèles.

Vclass1	ɛ e	*Cclass1*	b d g v
Vclass2	ɛ̃ ə œ ɔ	*Cclass2*	l r ɥ w j

En français beaucoup de quasi-homophones se différencient par la caractéristique ouvert/fermé de voyelles (e.g.: est /ɛ/, et /e/). Dans la parole courante cette distinction peut disparaître, l'identification correcte s'appuyant davantage sur le contexte que sur le signal acoustique.

Pour donner une indication de la complexité des différents dictionnaires nous indiquons dans la table 5 le rapport du nombre total de variantes dans le dictionnaire par le nombre total d'entrées lexicales. Les dictionnaires *Copt* admettent le plus de variantes. Les taux globalement plus élevées pour BREF s'expliquent simplement par un nombre plus élevé d'entrées lexicales plus longues. Pour les variantes parallèles peu de phonèmes peuvent être modifiés et les taux, relativement faibles, sont les plus forts pour la classe *Cclass2* des liquides et glissantes.

Table 5: Rapports $\frac{\#variantes}{\#entres}$ dans les dictionnaires références, séquentiels *Vopt*, *Copt*, parallèles *Vclass*, *Cclass*.

	MASK	*BREF*
Référence	1.1	1.2
Vopt	9.5	17.3
Copt	20.0	33.7
Vclass1	1.7	2.5
Vclass2	2.4	4.0
Cclass1	2.7	4.3
Cclass2	10.1	15.1

4. MESURE: LE TAUX DE *Variant2+*

Pour mesurer l'utilité des variantes lors de l'alignement automatique nous comptons la proportion de mots alignés avec les prononciations minoritaires. Cette mesure, appelée le taux de *Variant2+* [Add99], donne le pourcentage de mots alignés avec des variantes de rang de fréquence $r_\varphi \geq 2$. Ce taux donne une indication sur le besoin de variantes pour une meilleure modélisation acoustique oubien, de manière équivalente, sur la capacité d'une prononciation unique de rendre compte de toutes les occurrences de

ce mot. Le taux de *Variant2+* est définie dans les équations [1,2], où n désigne le mot de rang de fréquence n, $\#occ_n$ le nombre d'occurrences du mot n dans le corpus et $\#align_n^{r_\varphi=1}$ le nombre de mots alignés avec la variante majoritaire.

$$\%var2+_n = 100 \times (\#occ_n - \#align_n^{r_\varphi=1})/\#occ_n \quad [1]$$

$$\%Variant2+(n) = \frac{\sum_{i=1}^{n} var2+_i}{n} \quad [2]$$

La mesure $var2+_n$ ([1]) est spécifique au mot de rang n et le taux $Variant2+(n)$ ([2]) intègre tous les mots du rang 1 au rang n. Le taux de *Variant2+* **global** est le *%Variant2+*(**N**), avec **N** la taille du lexique.

Table 6: Exemple d'entrée lexicale n dans le dictionnaire *référence*, nombre d'occurrences dans BREF ($\#occ_n$), ltaux $var2+$ et détail des différentes variantes triées par rang de fréquence r_φ avec le nombre d'occurrences alignées ($\#align_n$).

entry n	$\#occ_n$	$var2+_n$	*phon.*	r_φ	$\#align_n$
les	21362	24%	le	1	16262
			lez	2	5100

Les figures 2-8 donnent le taux de *Variant2+* en fonction du rang de fréquence des mots. Dans chaque figure la courbe obtenue avec le dictionnaire de référence (du système de reconnaissance) est rajoutée, ce qui permet d'évaluer la proportion de mots mieux modélisés par les dictionnaires de variantes qu'avec les dictionnaires de reconnaissance.

5. CONDITIONS & RÉSULTATS

5.1. *Configurations d'alignement*

Des expériences d'alignement automatique ont été faites avec différents modèles acoustiques indépendants du contexte (CI: 36, 35) et dépendants du contexte (CD: 637, 594) pour MASK et BREF respectivement. Pour BREF un deuxième ensemble de 761 modèles CD a été utilisé.

La table 7 montre que le taux de *Variant2+* obtenus avec les dictionnaires *Vopt* et *Copt* décroît de manière significative en passant de modèles CI à des modèles CD. Cette même observation a pu être faite avec tous les dictionnaires de prononciation. Ce résultat, vérifié aussi sur l'anglais [Add99], montre que le besoin de variantes diminue si le nombre de modèles acoustiques augmente.

Table 7: Taux de *Variant2+* global pour différents modèles acoustiques et dictionnaires.

dictionnaires	mod. ac.	*MASK*	*BREF*
Vopt	*CI*	22.2	18.6
	CD	13.0	14.8
Copt	*CI*	27.0	21.0
	CD	14.5	16.2

5.2. *Styles de parole et types de variantes*

Dans les figures 2 et 3 le taux de *Variant2+* obtenu avec les dictionnaires *Vopt* sur la parole lue et spontanée est donné en fonction de la fréquence des mots. La figure 4 regroupe les mêmes informations pour les dictionnaires *Copt*.

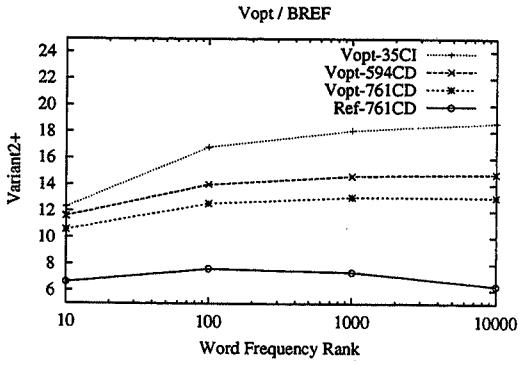


Figure 2: Taux de *Variant2+* en fonction du rang des mots pour la parole lue (BREF) avec le dictionnaire *Vopt* et différents modèles acoustiques.

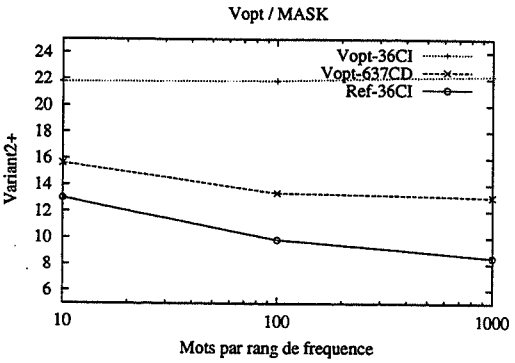


Figure 3: Taux de *Variant2+* en fonction du rang des mots pour la parole spontanée (MASK) avec le dictionnaire *Vopt* et différents modèles acoustiques.

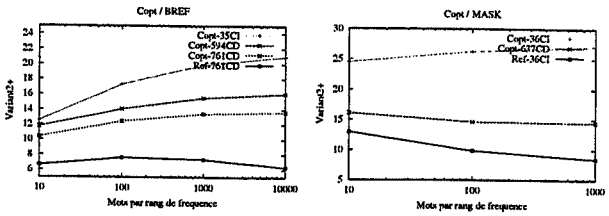


Figure 4: Taux de *Variant2+* en fonction du rang des mots pour BREF et MASK utilisant les dictionnaires *Copt* et différents modèles acoustiques.

Des courbes similaires sont données pour les dictionnaires *Vclass* and Cclass dans les figures 5- 8. Le taux de *Variant2+* est très élevé pour les mots fréquents en parole spontanée, la même chose n'est pas vrai pour la parole lue. Les dictionnaires *Vopt* et *Copt* admettent plus de variantes en parole spontanée qu'en parole lue, spécialement avec des modèles CI, mais les modèles CD permettent de réduire considérablement ce taux. Les modèles CD, appris avec les dictionnaires de référence, absorbent une part importante de ces variantes séquentielles, surtout pour la parole spontanée. On peut donc faire l'hypothèse que beaucoup de modèles de phone en contexte représentent des segments acoustiques différents d'un simple segment phonétique. L'analyse des résultats obtenus avec les dictionnaires de classes de consonnes (figures 5, 6) montrent que les modèles acoustiques des consonnes sont relativement précis. Pour la parole lue les taux de *Variant2+*

restent faibles même pour les modèles CI, ce qui n'est pas le cas pour la parole spontanée. Les modèles CD ramènent les courbes très près des courbes référence. Pour les classes de voyelles (figures 7 et 8) des variations plus importantes peuvent être mesurées. La classe *Vclass1* a un taux important de *Variant2+* avec une forte proportion de substitutions ɛ→e. La comparaison entre parole lue et spontanée (BREF et MASK) montre qu'avec les modèles acoustiques CI, la parole spontanée admet significativement plus de variantes que la parole lue et que l'emploi de modèles CD réduit toujours le taux de *Variant2+*: ceci est particulièrement vrai pour MASK où le vocabulaire limité fait que les modèles de phones contexte-dépendant deviennent vite mot-dépendent.

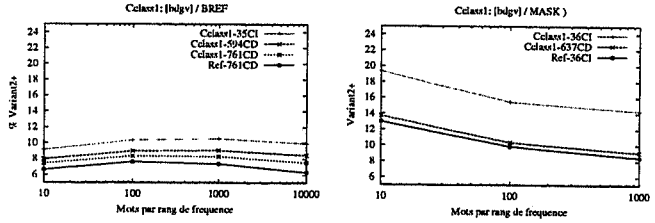


Figure 5: Taux de *Variant2+* en fonction du rang pour BREF and MASK avec les dictionnaires de *Cclass1* ([bdgv]).

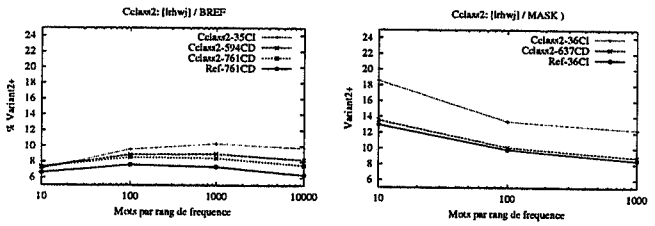


Figure 6: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Cclass2* ([lrɳwj]).

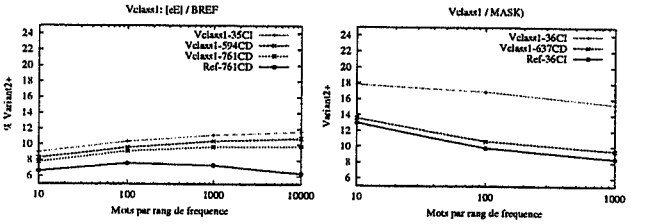


Figure 7: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Vclass1* ([eE]).

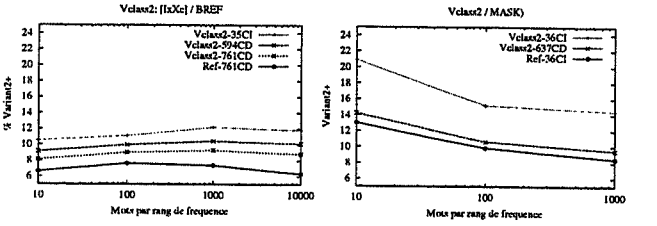


Figure 8: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Vclass2* ([ɛ̃ ə œ ɔ]).

À partir de ces mesures globales beaucoup d'investigations précises sont possibles. Par exemple nous avons examiné les clusters plosives-liquides en position finale de mot où un taux relativement fort de variation séquentielle peut être supposé. Dans le corpus BREF environ 25k mots sont concernés, 7k dans MASK. Des taux de *Variant2+* de 38% et de 51% sont obtenus avec des modèles CI et les dictionnaires *Copt* pour BREF et MASK respectivement.

6. Discussion & Perspectives

Des résultats comparatifs d'alignement utilisant différents ensembles de modèles acoustiques sur différents types de corpus avec des dictionnaires de pronnociations à taux de variantes élevés montrent que le besoin de variantes de prononciation dépend fortement de la précision des modèles acoustiques. Augmenter le nombre de modèles contexte-dépendants, couvrant progressivement plus de contextes diminue le besoin de variantes.

Les dictionnaires de variantes séquentielles montrent que si le choix est donné au système d'alignement, un pourcentage important des mots sont alignés avec un nombre de phonèmes différents du nombre prévu. En français la liaison et le schwa final optionnel sont des phénomènes bien connus permettant de générer un nombre variable de phonèmes par mot [Add99b]. Les modèles dépendants du contexte permettent d'absorber une partie de cette variabilité, particulièrement si ces modèles sont appris et utilisés sur un même vocabulaire de taille réduite (MASK, spontané). Les modèles deviennent ainsi plus spécifiques au mot et moins spécifiques au phonème.

Cette étude nous permet d'analyser le comportement global des modèles acoustiques de phones lors de l'alignement avec des dictionnaires très permissifs permettant de simuler localement la reconnaissance phonétique à un prix de calcul beaucoup plus faible. Cette approche permet de focaliser l'attention sur des problèmes précis qui peuvent se poser autant d'un point de vue d'ingénierie de la parole que d'un point de vue linguistique.

Bibliographie

[Spe99] **Speech Communication** "Special Issue on Pronunciation Variation Modeling", **29**, 1999.

[Lam96] L.Lamel, G.Adda, "On Designing Pronunciation Lexicons for LVCSR", *ICSLP'96*.

[Per98] G. Pérennou, L. Brieussel-Pousse, "Phonological Component in Automatic Speech Recognition", **ESCA-ETRW** Pronunciation Modeling for ASR, Rolduc, May 1998.

[Lus98] E. Fossler-Lussier, N. Morgan, "Effects of speaking rate and word frequency on conversational pronunciations", **ESCA-ETRW** Pronunciation Modeling for ASR, Rolduc, May 1998.

[Add99] M. Adda-Decker, L. Lamel, *"Pronunciation Variants Across Systems, Languages and Speaking Style,"*, Speech Communication, **29**, pp.83-99, 1999.

[Mok98] H. Mokbel, D. Jouvet, "Derivation of the optimal phonetic transcription set for a word from its acoustic realisations", **ESCA-ETRW** Pronunciation Modeling for ASR, May 1998.

[Add99b] M. Adda-Decker, P. Boula de Mareuil, L. Lamel, "Pronunciation variants in French: schwa & liaison", ICPhS-99, août 1999.

[Lam95] L. Lamel et al., "Development of Spoken Language Corpora for Travel Information", *EuroSpeech'95*.

[Lam91] L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *EuroSpeech'91*.

Parole et cognition

Traitement du langage parlé : resyllabation, liaison et enchaînement

Elsa Spinelli (1), Gareth Gaskell (2) & Fanny Meunier

MRC-CBU, 15 Chaucer Road, CB2 2EF,
Cambridge, England.

(1) maintenant au Max Planck Institute for Psycholinguistics
Wundtlaan 1, PB 310, 6500 AH Nijmegen, Pays Bas
Tel: (+31) (0)24 – 3521582
E-mail: elsa.spinelli@mpi.nl

(2) maintenant à University of York
Heslington, York YO10 5DD UK

ABSTRACT

French language is known to be a syllable timed language. Surprisingly, resyllabification occurs often in French due to either liaison *(un petit indien)*, enchainment (*un vase indien*) or elision (*l'indien*). In order to test whether this phenomenon is problematic for a segmentation strategy based on syllables, we carried out two cross-modal (auditory-visual) repetition priming experiments in which each target was preceded by either a liaison, an enchainement, a no liaison or a control condition. When the target is presented at the offset of the sentence prime (exp1) or 50 ms before the uniqueness point of the last word (exp2), we obtain a significant repetition priming effect in all conditions but no difference between the 3 critical conditions. The results suggest that there is no cost for the processing of linked words.

1 INTRODUCTION

Contrairement au langage écrit, la parole est un flux continu sans marque explicite de frontière de mot, la reconnaissance des mots passe donc, implicitement, par une étape d'extraction des mots parlés ([Lis87] ; [Cut90]). Afin d'isoler des mots dans la chaîne parlée, on serait en partie aidé par le fait que les frontières de mot coïncident avec celles des syllabes.

Des auteurs ont suggéré que le signal de parole est recodé en unités syllabiques et que ces unités peuvent être utilisées pour accéder au lexique. Dans une des études princeps sur le rôle de la syllabe, Mehler, Dommergues, Frauenfelder et Segui (1981) [Meh81] ont exploité la possibilité en Français, de sélectionner des mots qui ont la même séquence initiale de phonèmes mais dont les syllabes initiales sont différentes. C'est le cas de paires de mots comme « *balcon* » / « *balance* » qui partagent les trois phonèmes initiaux mais dont la syllabe initiale diffère. Les trois phonèmes /b/ + /a/ + /l/ constituent la première syllabe du mot « *balcon* » mais dépassent la frontière syllabique du mot « *balance* » qui est syllabé « *ba . lance* ». La tâche des sujets était de détecter des cibles telles que /ba/ ou /bal/ dans des paires de mots comme « *balcon* » ou « *balance* ».

Leurs résultats sont en faveur de l'hypothèse selon laquelle la syllabe constitue une unité de segmentation. Les temps de détection sont plus courts pour les cibles qui correspondent à la première syllabe du mot porteur que pour les cibles qui ne correspondent pas à la première syllabe. Ainsi, on détecte plus rapidement la cible « *bal* » dans « *balcon* » que dans « *balance* » et c'est l'inverse lorsqu'il s'agit de détecter « *ba* ».

Le problème du non alignement et de la resyllabation dans le traitement des mots :

Que se passe-t-il lorsque les débuts de mots ne correspondent pas aux débuts de syllabes ?

A l'aide, d'un paradigme d'amorçage sémantique intermodal (auditif-visuel), Vroomen & De gelder (1997) [Vro97] ont montré que la présentation auditive de *framboos* (*framboise* en Néerlandais) permet l'activation de *boos* (*en colère*). Si l'hypothèse lexicale *boos* est considérée lors de la présentation de *framboos*, c'est parce que *boos* est aligné avec le début d'une syllabe (et qui plus est, d'une syllabe forte). Ce n'est pas le cas de *vijn (vin)* qui est misaligné dans *zvijn (porc)*. Dans ce cas, les résultats montrent que l'hypothèse lexicale *vijn* n'est pas considérée. A l'aide d'une tâche de word spotting, Dumay, Banel, Frauenfelder & Content (1998) [Dum98] ont montré qu'il est plus facile de détecter le mot *lac* dans le nonmot *ZUN.LAC* (ou *lac* est aligné avec le début d'une syllabe) que dans le nonmot *ZU.GLAC* (ou *lac* ne correspond pas à un début de syllabe).

De manière analogue, Mc Queen, (1998) [McQ98] a montré qu'il est plus facile de détecter le mot *rok* (*jupe* en

Néerlandais) dans le nonmot *FIM.ROK* (condition alignée) que dans le nonmot *FI.DROK* (condition misalignée). Pris dans leur ensemble, ces études suggèrent qu'il y aurait un coût de traitement pour les mots qui sont non alignés avec les débuts des syllabes.

Vroomen & De gelder (1999) [Vro99] ont également montré que la resyllabation pouvait être problématique en Néerlandais. Leurs résultats montrent qu'il est plus difficile de détecter un phonème dans un mot resyllabé que dans un mot non resyllabé.

Dans cette optique, Yersin-Besson & Grosjean (1996) [Yer96] ont décrit deux processus phonologiques particulièrement intéressants en Français : celui de l'enchaînement et de la liaison enchaînée. L'enchaînement a lieu lorsqu'un mot contenant une consonne finale toujours réalisée est suivi par un mot commençant par une voyelle (ex : "*chaque avion*"). Le phénomène de liaison enchaînée consiste à réaliser une consonne finale lorsque le mot suivant commence par une voyelle mais pas lorsqu'il commence par une consonne (ex : "*petit avion*" mais "*petit cahier*" ; Encrevé, 1988[Enc88]). Ceci implique d'une part la prononciation d'un segment absent lorsque le mot est présenté en isolation (phonème latent) et d'autre part la resyllabation du mot suivant qui reçoit le phonème latent en segment initial. (ex: *un petit avion* sera resyllabé *un.pe.ti.ta.vion*).

Il résulte de ce phénomène, qu'après resyllabation, les frontières de mot ne coïncident plus avec les frontières de syllabe. Si la reconnaissance du mot comportant le phonème de liaison ne devrait pas poser de problème selon les modèles de reconnaissance, il en est autrement pour le mot suivant qui est réalisé avec la consonne latente en syllabe initiale. Ce phénomène soulève une question : dans quelle mesure la segmentation et, par là même, la reconnaissance des mots est influencée par la resyllabation due aux liaisons ?

Afin d'examiner cette problématique, nous avons effectué une première expérience d'amorçage intermodal de répétition où chaque mot cible (ex : ITALIEN) était présenté dans quatre conditions d'amorçage : une condition de liaison (*un généreux italien*-ITALIEN), une condition enchaînement (*un virtuose italien*-ITALIEN), une condition de non liaison (*un chapeau italien*-ITALIEN) et une condition contrôle (*un mystérieux organisme*-ITALIEN).

2 EXPERIENCE 1

2.1 *Méthode*

Sujets : Quarante sujets de langue maternelle française ont participé à l'expérience. Tous avaient une vision normale ou corrigée et n'ont reporté aucun trouble particulier de l'audition.

Stimuli

Mots expérimentaux :

Trente mots commençant par une voyelle (ex : ITALIEN) ont été sélectionnés dans la base de donnée TLF (trésor de la langue Française [Imb71]). A ces 30 mots cibles, étaient associées 120 phrases amorces. Trente d'entre elles constituaient la condition de liaison (*un généreux italien*). Trente autres constituaient la condition d'enchaînement (*un virtuose italien*), 30 autres constituaient la condition de non liaison (*un chapeau italien*) et les 30 dernières constituaient la condition contrôle (*un mystérieux organisme*).

Items de remplissage et essais attrapes :

Quarante cinq mots cibles ont été présentés dans une condition non reliée. Soixante quinze nonmots cibles ont été crées en respectant les contraintes phonotactiques du Français. Ils ont été présentés dans une condition non reliée. Par conséquent, la proportion de paires reliées est 29.33 %.

Passation de l'expérience

La passation s'effectue individuellement. Le sujet commence par entendre une phrase amorce présentée dans un casque (Sony CD550). A la fin de la phrase amorce, un item cible apparaît visuellement au centre de l'écran d'un ordinateur. Cet item reste affiché à l'écran pendant 480 ms.

La tâche du sujet est une tâche de décision lexicale : il doit décider le plus rapidement possible si l'item cible présenté à l'écran constitue un mot de la langue française ou pas. Il donne sa réponse en appuyant sur l'un des deux boutons mis à sa disposition (mot / nonmot). L'horloge de l'ordinateur est déclenchée à l'apparition de l'item cible et s'arrête à la réponse du sujet.

Afin que chaque sujet soit confronté à toutes les conditions expérimentales mais ne voie pas deux fois le même item cible, la répartition des mots cibles a été contrebalancée dans quatre listes. Les items sont divisés en quatre blocs (avec un nombre égal d'items de chaque condition expérimentale) et pour chaque sujet, ils sont présentés dans un nombre aléatoire au sein de chaque bloc.

Tableau 1 : Temps de réponse moyens (en Millisecondes), Ecarts types (entre parenthèses) et Taux d'erreurs (%) pour les cibles dans les 4 conditions d'amorçage.

	Liaison	**Enchaînement**	**Non liaison**	**Contrôle**
TR	540	532	526	585
ET	(104)	(95)	(81)	(72)
Erreurs	2.33 %	3.03 %	1.04 %	5.56 %

2.2 *Résultats*

Les réponses incorrectes (2.97 %) ont été enlevées. Quatre sujets ont été éliminés des analyses (1 à cause d'un

taux d'erreurs supérieur à 30 %, 1 à cause d'une moyenne de temps de réponse inférieur à 350 ms et 2 à cause d'une moyenne de temps de réponse supérieur à 700 ms)

Les données ont été transformées en moyennes harmoniques (1000/valeurs). Les analyses de variance effectuées sur ces données révèlent un effet global d'amorçage ($F1(3,96)=18.24$, $p<.001$; $F2(3,87)=21.02$, $p<.001$), ainsi qu'un effet significatif de répétition dans les conditions de liaison ($F1(1,32)=22.59$, $p<.001$; $F2(1,29)=44.49$,$p<.001$),d'enchaînement ($F1(1,32)=25.09$, $p<.001$; $F2(1,29)=52.95$, $p<.001$), et de non liaison ($F1(1,32)=35.14$, $p<.001$; $F2(1,29)=56.47$, $p<.001$). Par contre , il n'y a pas de différence entre ces trois conditions ($F1$ et $F2 < 1$ dans tous les cas).

Les résultats ont montré un effet de répétition par rapport à la condition contrôle dans tous les cas, mais pas de différence entre les 3 conditions critiques. Cependant, il est possible que la présentation du mot cible ait été trop tardive pour mettre en évidence une différence entre les 3 conditions. En effet, la différence attendue entre les 3 conditions qui nous intéressent est sensée refléter la résolution d'un conflit dû à une compétition transitoire entre des candidats lexicaux.

Etant donné que les amorces sont des énoncés présentés auditivement, donc par nature de façon séquentielle, il est probable qu'à la fin de la présentation du dernier mot amorce, ce conflit ait été résolu et les candidats en compétition ont retrouvé leur niveau de repos. Pour tester cette hypothèse, une expérience de gating a été réalisée (sur les stimuli réenregistrés de manière isolée) pour déterminer le Point d'Unicité des mots cibles. Une fois que nous avions les PU nous avons reconduit l'expérience en présentant le mot cible juste avant ce point.

3 EXPERIENCE 2

3.1 Méthode

Sujets : Quarante quatre sujets de langue maternelle française ont participé à l'expérience et ont été rémunérés £ 6.5 pour leur participation. Tous avaient une vision normale ou corrigée et n'ont reporté aucun trouble particulier de l'audition.

Stimuli et procédure

Les items sont les mêmes que ceux de l'expérience 1. Des marques ont été placées dans les phrases amorces 50 ms avant le point d'unicité du dernier mot (sur la base des résultats de l'expérience de gating). Ces marques constituaient le signal d'apparition des cibles visuelles.

3.2 Résultats

Les réponses incorrectes (1.65 %) ont été enlevées. Trois sujets ont été éliminés des analyses selon les mêmes critères que ceux appliqués pour l'expérience 1.

Tableau 2 : Temps de réponse moyens (en Millisecondes), Ecarts types (entre parenthèses) et Taux d'erreurs (%) pour les cibles dans les 4 conditions d'amorçage.

	Liaison	**Enchainement**	**Non liaison**	**Contrôle**
RT	560	560	562	602
ET	(74)	(81)	(76)	(72)
Erreurs	0.61 %	1.31 %	2.61 %	2.05 %

Les données ont été transformées en moyennes harmoniques (1000/valeurs).

Comme pour l'expérience 1, les analyses de variance effectuées sur ces données révèlent un effet global d'amorçage ($F1(3,111)=14.08$, $p<.001$; $F2(3,87)=8.86$, $p<.001$), ainsi qu'un effet significatif de répétition dans les conditions de liaison ($F1(1,37)=27.22$, $p<.001$; $F2(1,29)=26.31$, $p<.001$), d'enchaînement ($F1(1,37)=27.95$, $p<.001$; $F2(1,29)=16.18$, $p<.001$) et de non liaison ($F1(1,37)=24.08$, $p<.001$; $F2(1,29)=13.82$, $p<.001$). Par contre , il n'y a pas de différence entre ces trois conditions ($F1$ et $F2 < 1$ dans tous les cas).

4 DISCUSSION GENERALE

Les comparaisons critiques concernent les 3 conditions (liaison, enchaînement et non liaison). La condition non reliée fournissait un controle. Selon une hypothèse de segmentation syllabique, les temps de réponse devaient s'ordonner comme suit: la condition de non liaison devait fournir des temps plus courts que les conditions liaison et enchaînement puisque dans cette condition le mot ITALIEN n'est pas resyllabé, donc réalisé dans sa forme canonique. Des temps plus longs devaient être observés pour les deux conditions de resyllabation avec un désavantage pour la condition de liaison.

En effet, puisque le phonème « *z* » n'est pas toujours réalisé dans le mot « *généreux* », le système de traitement aura tendance à rattacher ce phonème au mot suivant (*italien*) qui ne contient pas ce phonème en position initiale. La reconnaissance du mot suivant nécessiterait donc une réanalyse. Ce problème est moindre pour la situation d'enchaînement dans la mesure ou le phonème final (ici « *z* » dans « *un virtuose italien* ») est toujours réalisé et fait partie de la représentation du mot « *virtuose* ».

Cependant, les résultats ne vont pas dans ce sens et il n'y a pas de différence entre les 3 conditions critiques. Ceci suggère que malgré la resyllabation (due soit à une liaison, soit à un enchaînement,) l'accès aux représentations lexicales n'est pas perturbé. Il n'y aurait donc pas de coût de traitement pour les mots resyllabés. Pourquoi ? Les phénomènes de liaison et d'enchaînement sont hautement systématiques en Français et il semble que les mots à voyelles initiales apparaissent plus fréquemment sous une

forme liée que dans leur forme canonique. Ceci pourrait être une piste pour expliquer la divergence entre nos résultats et ceux obtenus en Néerlandais par Vroomen & De Gelder (1999).

Quels peuvent être les mécanismes par lesquels le système résout le problème de la resyllabation ? Est ce que les mots à voyelles initiales peuvent être représentés avec une consonne initiale (latente) ? Ceci parait peu probable. En revanche, il est possible qu'il y ait dans le signal des indices permettant de différencier les consonnes de liaison des consonnes initiales, une consonne de liaison indiquerait alors clairement que la voyelle à venir est en position initiale de mot.

Ce point a été abordé par Durand (1936) [Dur36] dont les analyses acoustiques du signal de parole révèlent que la réalisation du phonème / t / prononcé dans un cas d'enchaînement (ex : "une petite orange") n'est pas la même que dans un cas de liaison enchaînée, donc prononcé en resyllabation (ex : "un petit orage"). Dejean de la Bâtie observe également une durée d'occlusion plus élevée et un VOT plus important dans la production d'un phonème en segment initial que dans la production du même phonème en situation de liaison.

Enfin il est également possible que le contexte lexical joue un rôle. Puisqu'un mot comme *petit* possède un */t/* sous-jacent pour la liaison, les hypothèses lexicales à voyelles initiales seraient privilégiées, ce qui ne serait pas le cas d'un mot comme *joli* qui n'a pas de consonnes sous-jacente pour la liaison. Si le contexte lexical joue un rôle dans le traitement de la liaison, alors le traitement de *avion* dans *grand avion* devrait être plus rapide que dans *vrai tavion* puisque *vrai* ne possède pas de /t/ sous-jacent pour la liaison. L'information lexicale permettrait de rattacher le /t/ à *grand* dans *grand avion* mais pas à *vrai* dans *vrai tavion* (recherche en cours).

Nos résultats ne nous permettent pas, pour l'instant, de conclure sur ces différentes hypothèses et des travaux supplémentaires sont nécessaires pour comprendre comment le système de traitement résout le problème de la resyllabation.

BIBLIOGRAPHIE

[Cut90] Cutler, A. & Butterfield, S. (1990). Syllabic lengthening as a word boundary cue. *In Proceedings of the third Australian International Conference on Speech Science and Technology,* pp. 324-328.

[Dur36] Durand, M. (1936). *Le genre grammatical en français parlé à Paris et en région parisienne.* Bibliothèque du Français Moderne, Paris.

[Lis87] Lisker, L. & Abramson, A. S. (1987). Phonetic validation of distinctive features : a test case in French. In Channon, Schockey (eds) *In honnour of Ilse Lehiste,* Dordrecht, Foris, pp.97-132.

[Meh81] Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior,* **20**, 298-305.

[Vro97] Vroomen, J. & De gelder, B. (1997). Activation of embedded words in spoken word recognition. Journal of Experimental Psychology : Human Perception and Performance, 1997, Vol 23 (3). 710-720.

[Dum98] Dumay, N., Banel, M.H., Frauenfelder, U.H. & Content, A. (1998).Le role de la syllabe : segmentation lexicale ou classification ? *Actes des XXII emes Journees d'Etude sur la Parole,* (pp.33-36), Martigny, Suisse.

[McQ98] Mc Queen, J.M. (1998). Segmentation of continuous speechusing phonotactics. *Journal of Memory and Language,* 39,21-46.

[Vro99] Vroomen, J. & De gelder, B. (1999). Lexical access of resyllabified words : evidence from phoneme monitoring. Memory & Cognition 1999, Vol 27 (3), 413-421.

[Yer96] Yersin-Besson, C. & Grosjean, F. (1996). L'effet de l'enchainement sur la reconnaissance des mots dans la parole continue. *L'année Psychologique, 96,* 9-30.

[Imb71] Imbs, P. (1971). *Trésor de la langue Française.* Dictionnaire des Fréquences. Paris: Klincksieck.

[Enc88] Encrevé, P. (1988). *La liaison avec et sans enchaînement,* Paris, Seuil.

Entraînement intensif des capacités phonologiques dans l'Aphasie Progressive Primaire : un modèle de plasticité cérébrale en pathologie neuro-dégénérative

Marianne Louis°, Michel Habib*°, Robert Espesser*, Virginie Daffaure°*
*& Albert Di Cristo**

- *Laboratoire Parole et Langage, Université de Provence, Aix-en-Provence,
- °Laboratoire Parole et Langage, équipe Parole et Dyslexie, Hôpital Nord, Marseille

Abstract

Three patients with typical Mesulam's syndrome (a non-fluent primary progressive aphasia) were trained daily with auditory exercices involving several aspects of phonological processing. The training material was designed to test the theory that increasing the duration of formant transitions should facilitate phonemic discrimination and awareness. Significant improvement on the tasks was demonstrated by all 3 patients. This generalized to other tasks such as non word repetition and reading. Such results (1) argue for intensive focused therapy in neuro-degenerative disorders, (2) may constitute a good model of brain plasticity in neuro-degenerative disorders in general, and (3) support theories of phonological processing emphasizing temporal features of the auditory signal.

1. Introduction

Actuellement, la neuropsychologie s'interroge sur l'efficacité et la justification d'une prise en charge orthophonique, souvent longue et relativement coûteuse, pour des patients atteints de pathologies cognitives liées à la sénescence. Ainsi, tout modèle permettant d'évaluer la plasticité du cerveau de patients âgés souffrant d'affections dégénératives peut apporter des arguments précieux à ce débat.

Nous avons choisi d'étudier, comme modèle de plasticité cérébrale, 3 cas d'aphasie progressive primaire (Syndrome de Mesulam), une pathologie qui, de manière particulièrement pertinente pour ce qui concerne notre présent propos, est caractérisée par une perte progressive et longtemps isolée du langage. Nous leur proposons de suivre un entraînement quotidien et régulier, constitué d'exercices de conscience phonologique, lesquels ont été préalablement testés et validés sur une population d'enfants en difficulté d'apprentissage du langage [Hab99].

Nous supposons que ce matériel permet de modifier les capacités du cerveau à traiter les informations auditives à caractère bref et en succession rapide, comme le sont les éléments de la parole comme par exemple, les transitions formantiques. Le système phonologique conditionne le processus audio-phonatoire de telle sorte qu'on est capable de discerner les sons (et donc de les reproduire) uniquement sur la base des oppositions indispensables au système linguistique de la langue maternelle. C'est la notion de crible phonologique, qui joue en quelque sorte un rôle de filtre pour les sons du langage.

La phonologie a été spécifiquement choisie parce que les manifestations du fonctionnement cognitif mises en évidence par l'activation cérébrale sont à présent bien connues et individualisées. Les techniques modernes d'imagerie cérébrale fonctionnelle, chez les individus normaux, ont montré que le stockage des constituants phonétiques de la parole dans la mémoire à court terme d'une part, et la boucle d'autorépétition subvocale d'autre part, pouvant impliquer une activité dans deux régions séparées de l'aire du langage de l'hémisphère gauche du cerveau [Pau93] et [Dém92]. Ces régions corticales, en association avec le cortex auditif temporal, seraient activées dans le traitement de la discrimination auditive, constituant un réseau probablement spécifiquement altéré dans les cas d'aphasie progressive non-fluente [Bél97] et [Cro98].

Il est concevable que cet entraînement spécifique des capacités phonologiques puisse modifier les mécanismes connus de l'activation cérébrale de cette fonction.

2. Patients et Méthode

2.1 Présentation des patients

Deux femmes, âgées respectivement de 64 et 71 ans, et un homme âgé de 77 ans, constituent le groupe des patients étudiés. Tous les trois présentent une aphasie progressive primaire non fluente avec, comme premiers symptômes, un manque du mot et un agrammatisme.

Si, la symptomatologie est celle d'une aphasie nettement prédominante, longtemps isolée, où les signes classiques de démence font défaut, l'évolution se caractérise

toujours, parallèlement à l'aggravation des troubles du langage, par l'apparition plus ou moins tardive d'autres troubles des fonctions mentales supérieures, en particulier mnésiques et comportementaux, aboutissant dans tous les cas à un syndrome démentiel terminal.

L'aphasie s'aggrave progressivement avec l'apparition d'une altération de la compréhension syntaxique et d'un déficit au niveau phonologique. Celui-ci se caractérise par des erreurs phonémiques en production, une dégradation de la mémoire de travail auditive, et des erreurs de discriminations phonétiques.

Un scanner cérébral et un examen débimétrique cérébral ont été réalisés pour tous les patients, et les résultats s'avèrent compatibles avec le diagnostic. Les troubles du langage ont été examinés avec la version française du Boston Diagnostic Aphasia examination [Maz71]. Le profil aphasiologique est, dans les trois cas, celui d'une aphasie mixte à prédominance expressive, touchant principalement la fluence et la répétition avec un nombre important de déformations phonémiques et une compréhension orale et écrite relativement préservée mais non intacte.

2.2 Présentation de la méthodologie

L'entraînement comprend des séances de 15 à 20 minutes par jour, effectuées soit avec l'aide d'un orthophoniste, soit sous son contrôle, par un membre de la famille. Nous avons choisi des exercices purement mécaniques de discrimination auditive ou de reproduction de logatomes, les séances de « bain sonore » qui, par une mise en condition sur la base des éléments fréquentiels, séquentiels et suprasegmentaux facilitent l'intégration d'un système phonologique nouveau. Les exercices sont constitués principalement d'épreuves de détection d'intrus phonologiques, de comptage syllabique, de segmentation phonémique et de discrimination (table 1).

Table 1 : Exemple d'un type d'exercices

Consigne : Quels sont les deux mots qui se terminent par le même « son » (après quelques exemples d'apprentissage les sujets parviennent facilement à comprendre la notion de phonème par opposition à la notion de syllabe).

Exercice 1 :

bûche	douche	cage
beige	louche	juge
ânesse	douze	bosse
chaise	danse	rose

En accord avec la théorie temporelle du traitement phonémique [Tal96] et [Mer96] et dans le but de faciliter et, éventuellement entraîner les aptitudes supposées déficitaires de la perception des éléments rapides, nous avons modifié le signal de parole. Celui-ci a été globalement ralenti jusqu'à 166%, avec une augmentation proportionnelle d'énergie au niveau des portions instables, présumées correspondre aux éléments les plus rapides de la parole que sont les transitions formantiques. Les exercices ont été enregistrés sur un CD audio et diffusés par un casque de haute fidélité.

Le groupe a reçu un entraînement d'une durée de 42 jours dans des conditions similaires. Pour chaque jour, les réponses des sujets étaient notées sur une grille appropriée.

Un des patients a pu bénéficier de trois entraînements successifs. La première série (période n°1), commune au groupe, a duré 42 jours avec un entraînement réalisé en parole modifiée. La seconde série (période n°2), a duré 30 jours avec un entraînement basé sur de la parole normale. Une troisième série (période n°3) qui a également duré 30 jours avec un entraînement à nouveau conçu à partir de parole modifiée. Les exercices d'entraînement de la conscience phonologique sont similaires dans les 3 séries.

3. Résultats

Les performances quotidiennes de chacun des trois patients sont exposées dans la figure 1. Chaque point représente la performance moyenne de chaque patient sur les exercices effectués au cours d'une session. Une représentation linéaire de la progression moyenne des trois patients est présentée séparément sur le graphique, démontrant une amélioration globale et graduelle des performances au cours des 42 jours d'entraînement.

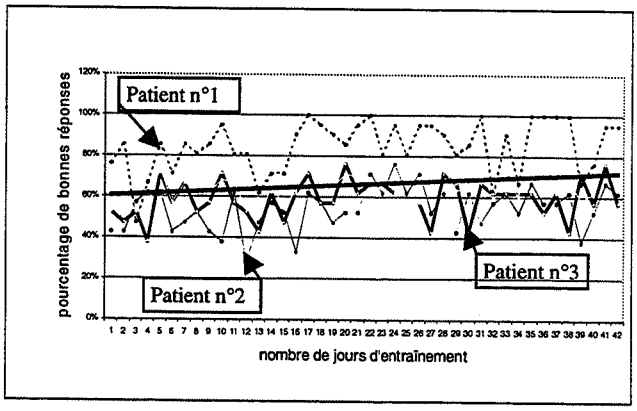


Figure1 : performance individuelle par jour d'entraînement avec des exercices de conscience phonologique en parole modifiée. La ligne noire correspond à la représentation linéaire de la progression de la moyenne des trois sujets.

L'examen des profils du BDAE souligne une amélioration significative de la fluence chez un patient, de la compréhension écrite chez un autre, de la répétition dans deux cas et de la lecture également dans deux cas. Il n'y a ni amélioration en dénomination ni en compréhension orale. Le nombre de paraphasies phonémiques s'est réduit

chez un patient alors qu'il est resté inchangé chez les deux autres.

Il nous semble important de souligner que dans les tâches spécifiques de lecture et de répétition nous avons trouvé une amélioration significative des performances, avant et après entraînement en parole ralentie (Mann-Whitney U, respectivement p = 0.007 et p = 0.002), (figure 2).

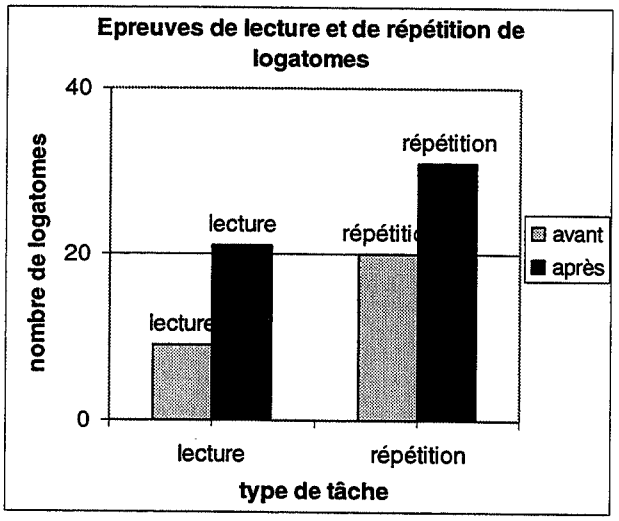


Figure2 : Tâches de lecture et de répétition de non-mots avant et après l'entraînement d'exercices de conscience phonologique en parole modifiée.

En dépit de l'aggravation globale des capacités langagières au cours de la période d'entraînement, certains aspects de l'évaluation neurolinguistique sont restés stables alors que d'autres se sont améliorés de manière significative. Ces aspects étant hypothétiquement liés au traitement phonologique, nous pourrions conclure que cet entraînement spécifique a été au moins efficace sur un sous-ensemble des fonctions langagières de ces patients.

Dans le but de confirmer cette tendance, le patient 1 a donc reçu deux autres séries d'exercices, la première (période n°2) en parole normale, non modifiée et la dernière (période 3), avec à nouveau un entraînement en parole modifiée. La comparaison de la représentation linéaire de la performance moyenne de chaque jour sur les exercices phonologiques ne montre pas d'amélioration dans la condition « parole normale » de la période n°2. En revanche, nous remarquons une progression des performances dans les conditions « parole ralentie » des périodes n°1 et n°3 (figure 3). Une analyse de type ANOVA à mesures répétées nous a permis de mettre en évidence une différence significative entre les deux conditions (F = 2.995 ; p = 0.0021).

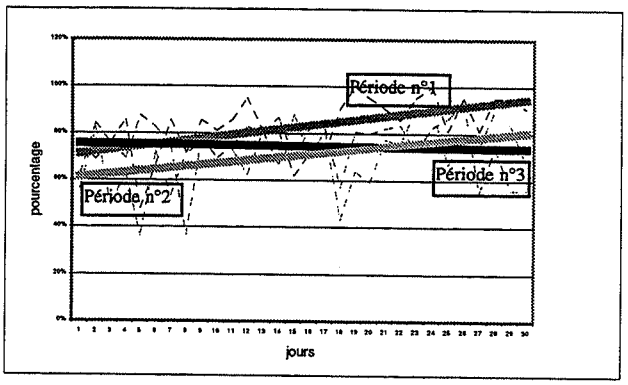


Figure3 : Représentation linéaire de la progression des performances lors de 3 phases successives de l'entraînement chez le sujet 1. La 1ère et la 3ème période de 30 jours (1 et 3) sont réalisées avec de la parole acoustiquement modifiée. Seule, la période 2, en parole normale, n'entraîne pas d'amélioration significative.

4. Discussion

Ce travail, préliminaire et exploratoire, nous semble posséder des implications théoriques et pratiques intéressantes.

D'un point de vue théorique, les résultats suggèrent l'efficacité potentielle sur le processus dégénératif (ou malgré le processus dégénératif) d'un entraînement intensif, focalisé sur un domaine très précis des fonctions cognitives perturbées. En cela, ils prouvent que le cerveau de patients souffrant de telles affections peut encore se réorganiser dans le sens d'une récupération partielle de modules déficients.

Ces résultats tendent donc à justifier que les malades atteints d'une pathologie neuro-dégénérative impliquant des altérations du langage, méritent de recevoir une intervention active. En effet, malgré la perte progressive et inéluctable tant structurelle que fonctionnelle qui caractérise ce groupe d'affections, un bénéfice significatif d'une intervention spécifique a pu être constaté, démontrant par là même, une possibilité de récupération même partielle de la fonction.

Dans un contexte différent, l'aphasie progressive dégénérative apparaît donc comme un terrain d'études potentiellement pertinent face au débat actuel opposant les théories linguistiques aux théories perceptives [Mod97], [Lib96].

Nos résultats confortent l'hypothèse selon laquelle les caractéristiques temporelles du signal de parole participent au processus de résolution, par le cerveau, de difficultés phonologiques. En effet, seul l'entraînement incluant de la parole modifiée temporellement a été spécialement efficace sur les processus langagiers relatifs à la phonologie.

Enfin, ces résultats incitent à poursuivre les études réalisées selon ce schéma, c'est à dire concernant une pathologie spécifique dont les perturbations sont relativement focalisées, sous forme d'un entraînement intensif et orienté spécifiquement vers un domaine précis des fonctions du langage.

5. Bibliographie

[Bél97] Béland R. (1997) "Principled syllabic dissolution in primary progressive aphasia case", Aphasiology, Vol. 11 (12), pp. 1171-1196.

[Cro98] Croot .K (1998), "Single word production in nonfluent progressive aphasia", Brain and Language, Vol. 61, pp. 226-273.

[Dém92] Démonet J.F. (1992) " The anatomy of phonological and semantic processing in normal subjects", Brain, Vol. 115, pp. 1753-1768.

[Hab99] Habib M. (1999) "Training dyslexics with acoustically modified speech : evidence of improved phonological performance" (abstract), Brain and cognition, Vol. 40 (1), pp. 143-146.

[Lib96] Liberman A.H. (1996) "Speech : a special code", M.I.T. Press, Cambridge, Mass.

[Mer96] Merzenich M.M. (1996), "Temporal processing deficits of language-learning impaired children ameliorated by training", Science, Vol. 271, pp. 77-80.

[Mod97] Mody M. (1997), "Speech perception deficits in poor readers : auditory processing or phonological coding ?", Journal Exp. Child Psychol., Vol. 64 (2), pp. 199-231.

[Pau93] Paulesu E. (1993), "The neural correlates of the verbal component of working memory", Nature, Vol. 362, pp. 342-344.

[Tal96] Tallal P. (1996), "Language comprehension in language-learning impaired children improved with acoustically modified speech", Science, Vol. 271, pp. 81-84.

Tolérance aux variations phonétiques dans l'accès au lexique : pourquoi "dlaïeul" est-il mieux toléré que "droseille" ?

Pierre A. Hallé et Juan Segui

Laboratoire de Psychologie Expérimentale, CNRS-Paris 5
Centre Universitaire de Boulogne - Boulogne-Billancourt, France
Tél.: ++33 (0)155 20 59 34
Mél: halle@psycho.univ-paris5.fr

ABSTRACT

Two experiments, using lexical decision and cross-modal repetition priming, investigated the role of the phonotactically based perceptual assimilation of /dl, tl/ as /gl, kl/ in the mapping of the speech signal onto lexical representations. Lexical activation was only slightly weakened when words such as *glaïeul* were altered into *dlaïeul.* By contrast, the same velar-to-dental single-feature alteration greatly reduced lexical activation for words such as *groseille* altered into *droseille.* These findings suggest that the mapping from speech signal to lexical representations does not proceed directly from a time-distributed array of phonetic features to a lexical code, but involves intermediary stages of perceptual integration into phonological units such as syllable onsets or, perhaps, larger units.

1. INTRODUCTION

L'accès au lexique dans la modalité auditive pose deux problèmes importants : la nature des représentations lexicales, et celle des processus d'accès.

Pour qu'un mot soit reconnu, l'adéquation entre forme physique d'entrée et forme canonique peut ne pas être parfaite. Par exemple, *shigarette* est encore reconnu comme *cigarette* [Nor82]. On peut interpréter cette relative souplesse du sytème de traitement de plusieurs façons. La représentation d'un mot pourrait être détaillée mais multiple, correspondant aux "traces" mnésiques de tous les exemplaires entendus ([Gol98], [Nos97]). Elle pourrait être de nature abstraite ou prototypique, mais rester relativement sous-spécifiée ([Lah91], [Mar94], [Ste86]). Enfin, les processus d'accès eux-mêmes doivent permettre de s'accommoder de certaines variations. Par exemple, [hæm] dans le contexte "Hand me the book" est interprété comme un exemplaire de *hand* et non de *ham* bien que la forme phonétique indique le contraire. S'il est impératif que les variations phonologiquement motivées soient tolérées voire exploitées [Gas96], les variations non régulières, accidentelles, ont moins de raison de l'être. Mais tant que ces variations n'excèdent pas un ou deux traits phonétiques, elles sont tolérées par le système [Con93]. Les modèles actuels incorporent cet aspect de l'accès au lexique et s'accommodent de variations limitées du signal d'entrée (Cohorte II [Mar87], TRACE [McC86], Merge [Nor00]). Ils s'opposent cependant sur le niveau auquel opèrent les processus d'appariement.

L'accès direct au lexique suppose que les représentations lexicales soient "directement" activées par une matrice de traits phonétiques distribués dans le temps, sans l'intervention d'unités sublexicales d'intégration de type phonème ou syllabe ([Mar94], [Ste86]). D'autres modèles supposent que le signal d'entrée soit analysé et intégré en unités successives, qui sont celles mêmes qui composent les représentations lexicales (Merge [Nor00]). Les données empiriques ne permettent guère de trancher (mais voir [Mar94] pour une défense de l'accès direct).

Dans cette étude, nous utilisons un cas d'assimilation perceptive — /dl, tl/ perçus comme /gl, kl/ en position initiale [Hal98] — pour apporter un nouvel éclairage sur cette question. Ce cas d'assimilation a été démontré avec des non-mots (comme *dlopta* ou *tlabod*) : il opère sans doute à un niveau sublexical d'intégration perceptive et en tout cas à un niveau d'intégration plus large que le phonème et a fortiori le trait. Si l'accès au lexique est direct, des formes comme *dlaïeul* et *droseille*, qui ne diffèrent toutes deux des mots *glaïeul* et *groseille* que par le changement de trait [+vélaire] en [+dental] pour la consonne initiale, devraient activer ces mots d'une manière comparable. Par contre, si l'accès au lexique passe par une phase d'intégration perceptive qui permet de construire /gl/ à partir de /d/ + /l/, et plus trivialement, /dr/ à partir de /d/ + /r/, la forme *dlaïeul* devrait activer davantage le mot *glaïeul* que *droseille* n'active *groseille.* En effet, à l'issue de cette phase d'intégration, les formes en /dl, tl/ deviennent virtuellement identiques aux mots sous-jacents en /gl, kl/, alors que les formes en /dr, tr/ restent déviantes. Ces prédictions sont testées en utilisant deux paradigmes expérimentaux : décision lexicale et amorçage inter-modal qui, indirectement ou directement, permettent de mesurer l'activation lexicale.

2. DÉCISION LEXICALE

Dans cette expérience, des mots intacts comme *glaïeul* et *groseille* sont comparés à des formes (non lexicales) comme *dlaïeul* et *droseille* obtenues par le changement de /g/ en /d/ ou de /k/ en /t/.

2.1 Méthode

Les items test étaient 48 mots commençant par /gl/, /kl/, /gr/, ou /kr/ (équilibrés pour la fréquence lexicale et la durée physique) et les non-mots dérivés par changement de /g/ en /d/ ou de /k/ en /t/. Ce matériel était scindé en deux listes expérimentales de façon à ce que les deux versions d'un même mot (intacte vs. altérée) apparaissent dans des listes différentes. Chaque liste comprenait ainsi 24 des 48 mots test et 24 formes altérées dérivées des 24 autres mots test, complétés par 24 mots et 48 non-mots de remplissage (soit en tout 120 items par liste). Abstraction faite des non-mots dérivés des mots test, dont le statut était a priori incertain, il y avait donc autant de mots que de non-mots (48).

La moitié des sujets étaient affectés à l'une des deux listes expérimentales, et l'autre moitié à l'autre liste de sorte à contrebalancer le type d'item (intact vs. altéré) entre les deux listes. Pour chaque sujet, la liste expérimentale était précédée d'une courte liste d'entraînement de dix mots et dix non-mots. Les items étaient présentés dans un ordre aléatoire. Pour chaque item, les sujets devaient répondre en appuyant sur l'un des deux boutons réponse étiquetés "mot" et "non-mot", le plus rapidement et le plus exactement possible. Les temps de réponse (TR) étaient mesurés à partir de l'onset acoustique des items.

Vingt étudiants de Paris V ont participé à l'expérience. Les données de quatre autres sujets n'ont pas été retenues (plus de 50% d'erreurs sur les items "intacts").

2.2 Résultats et discussion

Les résultats sont résumés dans la table 1. Les réponses "mot" étaient plus fréquentes pour les mots intacts en /gl, kl/ ou /gr, kr/ (93%) que pour les formes altérées (44%), mais parmi ces dernières, les formes en /dl, tl/ comme *dlaïeul* étaient jugées comme mot bien plus souvent que les formes en /dr, tr/ (68% vs. 20%), une différence significative dans les analyses par items et par sujets au niveau $p < .0001$. Les latences des réponses "mot" étaient plus longues dans le cas des formes altérées d'environ 70 ms pour les items en /dl, tl/ ($p < .005$). (Les TR pour les items en /dr, tr/ correspondent à trop peu de réponses pour être interprétables.)

Table 1: Pourcentages de réponses "mot" et temps de réponse correspondants.

	Mots intacts		Mots altérés	
exemples	*glaïeul*	*groseille*	*dlaïeul*	*droseille*
réponses "mot"	92.9%	93.8%	68.3%	19.6%
TR	814 ms	852 ms	884 ms	1015 ms

Cette expérience très simple montre que le changement vélaire-dental est beaucoup plus handicapant pour des mots comme *groseille* que pour des mots comme *glaïeul*. Il s'agit pourtant du même changement dans les deux cas. En termes de matrice de traits phonétiques, l'adéquation entre signal d'entrée et code lexical devrait donc être affaiblie d'une façon comparable. Pourtant les résultats suggèrent que *dlaïeul* est bien plus proche de *glaïeul* que *droseille* de *groseille*. Ce ne peut être que parce que l'assimilation dental-vélaire joue pour /dl, tl/, pas pour /dr, tr/. Mais joue-t-elle de façon on-line, en combinaison avec les processus d'appariement entre forme d'entrée et forme du lexique mental, ou bien n'intervient-elle qu'après coup, comme un mécanisme de récupération post-perceptuel ? C'est ce que pourraient suggérer les TR plus longs pour *dlaïeul* que pour *glaïeul*. (Mais ceci peut tout aussi bien refléter une surcharge de traitement *on-line*.) D'une façon plus générale, la tâche de décision lexicale requiert une décision explicite sur les stimuli auditifs présentés. Les résultats peuvent partiellement refléter des stratégies décisionnelles conscientes plutôt que des mécanismes on-line. Une tâche plus indirecte, l'amorçage inter-modal, permet de "mesurer" l'activation des mots susceptibles d'être activés lors de la présentation de stimuli auditifs sans que les sujets n'aient à répondre à ces stimuli. Elle est utilisée dans la deuxième expérience.

3. AMORÇAGE INTER-MODAL

Dans cette expérience, un mot présenté visuellement, par exemple GLAÏEUL, peut être précédé par *glaïeul* (répétition), *dlaïeul* (altération), ou par un mot non relié (comme *fauteuil*), présentés auditivement. De même, GROSEILLE peut être précédé par *groseille*, *droseille*, ou par un mot non relié. Ce montage expérimental doit permettre de savoir dans quelle mesure les formes altérées activent ou non les mots sous-jacents, selon qu'il y a assimilation perceptive ou pas.

3.1 Méthode

Les items cibles critiques étaient les 48 mots en /gl, kl/, et /gr, kr/ utilisés précédemment. Ces mots cibles pouvaient apparaître dans trois conditions d'amorçage: précédés par la forme auditive du même mot (*groseille* GROSEILLE), par une forme altérée par changement de l'initiale vélaire en dentale (*droseille* GROSEILLE), ou par un mot non relié (*fourchette* GROSEILLE). Pour que chaque mot cible ne soit vu qu'une seule fois par les sujets, les essais étaient répartis en trois listes de sorte à ce que chaque cible apparaisse une seule fois dans chaque liste mais dans trois conditions d'amorçage différentes selon les listes. Ces listes comportaient donc 48 essais critiques (mots cibles critiques en /gl, kl/ et /gr, kr/). Elles comportaient en outre 144 essais de remplissage destinés à équilibrer les proportions de mots et de non-mots en cible et en amorce et à éviter que le statut lexical des cibles soit prédictible. Ainsi, dans chaque liste, les cibles mots étaient précédées par des mots aussi souvent que des non-mots et de même pour les cibles non-mots ; chaque liste comprenait 24 non-mots cibles en /gl, kl/ et 24 en /gr, kr/ pour éviter que les suites graphémiques correspondantes n'induisent un biais de réponse "mot". Les amorces en /dl, tl/, /gl, kl/, /dr, tr/, et /gr, kr/ étaient suivies par de mots (32) ou des non-mots (48) toujours par souci d'éviter un

biais vers les réponses "mot". En dehors de ce luxe de précautions, l'essentiel reste que les cibles en /gr, kr/ et celles en /gl, kl/ étaient à armes égales dans les trois conditions d'amorçage.

Trois groupes de sujets étaient affectés à l'une des trois listes de sorte à contrebalancer cible et type d'amorçage (répétition, altération, et non-relié). Pour chaque sujet, la liste expérimentale était précédée d'une courte liste d'entraînement de 12 essais avec pour cibles six mots et six non-mots, précédés par une amorce "identique" ou non-reliée. Les essais se succédaient dans un ordre aléatoire. A chacun d'eux, les sujets devaient donner une réponse de décision lexicale *sur la cible visuelle*, le plus rapidement et le plus exactement possible. Les temps de réponse (TR) étaient mesurés à partir de l'apparition de la cible sur l'écran d'ordinateur.

Trente étudiants de Paris V ont participé à l'expérience. Les données de six autres sujets, qui avaient fait plus de 30% de décisions lexicales erronées sur l'ensemble des mots cibles, n'ont pas été retenues.

3.2 Résultats et discussion

Les résultats bruts sont résumés dans la table 2. Les conditions d'amorçage de répétition et d'altération entraînaient toutes deux une facilitation par rapport à la condition contrôle non-reliée, tant pour les TR que pour les taux d'erreurs. La facilitation en termes de diminution de TR est illustrée dans la figure 1. Dans la condition de répétition, elle était identique pour les mots en /gl, kl/ et ceux en /gr, kr/ (137 ms et 135 ms, respectivement). Par contre, dans la condition d'altération, la facilitation était bien plus forte pour les mots en /gl, kl/ (112 ms) que pour ceux en /gr, kr/ (52 ms) (par sujets: $F_1(1,27) = 14.4$, $p = .0008$; par items: $F_2(1,46) = 5.5$, $p = .022$).

La forme *dlaïeul* activait donc le mot "glaïeul" presque autant que la forme *glaïeul* (facilitation: 112 vs. 137 ms) tandis que la forme *droseille* activait le mot "groseille" bien moins que la forme *groseille* (facilitation: 52 vs. 135 ms). Ce pattern de résultats était reflété par une interaction significative entre condition d'amorçage et type de cible ($F_1(1,27) = 31.4$, $p < .0001$; $F_2(1,46) = 7.7$, $p < .01$), illustrée en figure 1.

Table 2: Décision lexicale sur les mots cibles critiques : TR en ms et pourcentages d'erreurs (entre parenthèses).

	Condition d'amorçage		
Mots Cibles	Répétition	Altération	Non-relié
GLAÏEUL	471 (3.0%)	496 (7.6%)	608 (12.2%)
GROSEILLE	469 (3.3%)	552 (6.8%)	604 (10.0%)

Les résultats de cette expérience d'amorçage inter-modal sont donc en accord avec ceux de l'expérience précédente. Ils ne peuvent cette fois-ci s'expliquer par une réanalyse a posteriori des stimuli auditifs. Comme la décision des sujets portait sur la cible visuelle et non l'amorce, l'activation lexicale induite par l'amorce ne pouvait guère être contrôlée par les sujets. Autrement dit, nous pouvons légitimement supposer que le mécanisme d'activation de la cible par l'amorce est automatique et opère en temps réel (cf. [Con93], [Con94], [Gas96], [Mar96]).

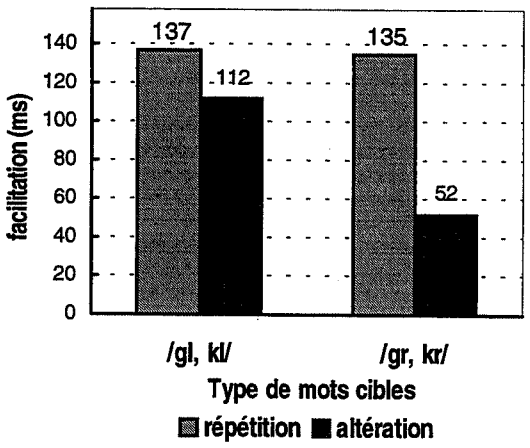


Figure 1: Facilitation en fonction de la condition d'amorçage et du type de mot cible.

La tâche de décision lexicale en auditif permettait d'inférer indirectement l'activation d'un mot tel que *glaïeul* par la forme *dlaïeul* : cette forme était jugée comme un mot, mais quel mot ? Le paradigme d'amorçage, quant à lui, permet de savoir que *dlaïeul* active bien *glaïeul* et pas d'autres mots. Ce paradigme est en outre plus sensible que la simple décision lexicale en ce qu'il permet de montrer une activation "résiduelle" (pour employer le terme de Gaskell et Marslen-Wilson [Gas96]) mais néammoins substantielle pour des formes telles que *droseille*. Ceci est en accord avec les résultats de Connine et al. [Con93] qui montrent qu'un changement d'un ou deux traits de la consonne initiale ou médiale laisse subsister une telle activation résiduelle.

A. Discussion générale

Les résultats de cette étude ont un double intérêt. D'une part, ils confirment pour des mots l'existence d'un phénomène d'assimilation perceptive, motivé par des contraintes phonotactiques, découvert initialement pour des non-mots [Hal98]. En effet, l'avantage de *dlaïeul* sur *droseille* ne s'explique aisément que si l'on suppose que /d/ + /l/ est intégré en /gl/ tandis que /d/ + /r/ est intégré en /dr/. (Un autre cas d'assimilation induite par des contraintes phonotactiques a été récemment rapporté pour des sujets japonais [Dup99]). D'autre part, la différence d'activation induite par les deux types de formes altérées /dl, tl/ et /dr, tr/ suggère que les formes lexicales ne sont pas directement accédées par une matrice de traits, ni a fortiori par une description acoustique encore plus détaillée. Elles sont accédées par l'intermédiaire d'unités d'intégration infra-lexicales qui restent à définir. Elles sont au minimum de la taille du cluster, ou plus généralement de l'onset syllabique, mais pourraient également correspondre à la syllabe ou à l'ensemble de gestes articulatoires qui sous-tendent la production d'une syllabe. Nos résultats mettent ainsi en difficulté les modèles qui

proposent des représentations des formes lexicales détaillées en deçà de l'onset syllabique, c'est à dire non seulement les modèles d'accès direct ([Kla89], [Mar94]), mais aussi ceux selon lesquels les représentations sont les "traces épisodiques" des multiples exemplaires qui constituent l'expérience qu'un auditeur a des mots qu'il reconnaît ([Gol98], [Nos97]).

Les représentations lexicales sont relativement rigides puisque, en l'absence d'assimilation perceptive, le changement d'un seul trait phonétique de la consonne initiale fait chuter l'activation d'un mot, telle qu'indexée par la facilitation observée dans une situation d'amorçage inter-modal. On peut ainsi affirmer que *droseille* active le mot "groseille" deux fois moins que *groseille*. Cette activation "résiduelle" est néanmoins réelle puisque la facilitation observée reste substantielle (environ 50 ms). Il y a donc une certaine tolérance aux variation phonétiques, même dans une situation hors contexte.

L'utilité de cette tolérance aux variations est évidente, étant donné la variabilité des énoncés de parole. Elle doit cependant être modulée par la viabilité des variations en fonction du contexte phonologique ([Gas96]). Nos données ajoutent à ce tableau que les assimilations perceptives motivées par des contraintes phonotactiques, bien qu'elles opèrent largement à un niveau non conscient, aident aussi à l'interprétation des énoncés déviants en sorte qu'ils fassent sens.

BIBLIOGRAPHIE

[Con93] Connine, C., Blasko, D., & Titone, D. (1993) "Do the beginnings of spoken words have a special status in auditory word recognition?", Journal of Memory and Language, Vol. 32, pp. 193-210

[Con94] Connine, C., Blasko, D., & Wang, J. (1994) "Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context", Perception and Psychophysics, Vol. 56, pp. 624-636.

[Dup99] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999) "Epenthetic vowels in Japanese: A perceptual illusion?", Journal of Experimental Psychology: Human Perception and Performance, Vol. 25, pp. 1568-1578.

[Gas96] Gaskell, G., & Marslen-Wilson, W. (1996) "Phonological variation and inference in lexical access", Journal of Experimental Psychology: Human Perception and Performance, Vol. 22, pp. 144-158.

[Gol98] Goldinger, S.D. (1998) "Echoes of echoes? An episodic theory of lexical access", Psychological Review, Vol. 105, pp. 251-279.

[Hal98] Hallé, P., Segui, J., Frauenfelder, U., & Meunier, C. (1998) "Processing illegal consonant clusters: A case of perceptual assimilation?", Journal of Experimental Psychology: Human Perception and Performance, Vol. 24, pp. 592-608.

[Kla89] Klatt, D.H. (1989) "Review of selected models of speech perception", in W. Marslen-Wilson (Ed.) *Lexical representation and process* (pp. 169-226), MIT Press.

[Lah91] Lahiri, A., & Marslen-Wilson, W. (1991) "The mental representation of lexical form: A phonological approach to the mental lexicon", Cognition, Vol. 38, pp. 245-294.

[Mar87] Marslen-Wilson, W. (1987) "Functional parallelism in spoken word recognition", Cognition, Vol. 25, pp. 71-102.

[Mar94] Marslen-Wilson, W., & Warren, P. (1994) "Levels of perceptual representation and process in lexical access: Words, phonemes, and features", Psychological Review, Vol. 101, pp. 653-675.

[Mar96] Marslen-Wilson, W., Moss, H., & van Halen, S. (1996) "Perceptual distance and competition in lexical access", Journal of Experimental Psychology: Human Perception and Performance, Vol. 22, pp. 1376-1392.

[McC86] McClelland, J.L., & Elman, J.L. (1986) " The TRACE model of speech perception", Cognitive Psychology, Vol. 18, pp. 1-86.

[Nor00] Norris, D., McQueen, J., & Cutler, A. (2000) "Merging information in speech recognition: Feedback is never necessary", Behavioral and Brain Sciences, Vol. 23, pp. Xxx-xxx.

[Nor82] Norris, D. (1982) "Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler", Cognition, Vol. 62, pp. 714-719.

[Nos97] Nosofsky, R., & Palmeri, T. (1997) "An exemplar-based random walk model of speeded classification", Psychological Review, Vol. 104, pp. 266-300.

[Ste86] Stevens, K.N. (1986), "Models of phonetic recognition II: A feature-based model of speech recognition", 12th International Congress on Acoustics, Montreal satellite Symposium on speech recognition, pp. 67-68.

La difficulté de l'accès lexical aux noms de personnes: évaluation au moyen d'une tâche de dénomination à partir de photographies

Muriel EVRARD
Laboratoire Jacques Lordat, Maison de la Recherche, Université de Toulouse-Le Mirail
5, allées Antonio Machado
31058 TOULOUSE CEDEX
tél.: 05.61.50.35.96; e-mail: Muriel.Evrard@univ-tlse2.fr

Abstract

This article presents data from a sample of 98 « normal » adults who had to name orally digitized pictures of objects (production of common nouns) and celebrities (production of names of people). In accordance with certain neuro-psychological studies of brain-damaged patients and findings from diary investigations, this laboratory study suggests that proper names are harder to recall than common nouns. Indeed, faces were named less quickly and, more importantly, induced more tip-of-the-tongue states than objects. The results are discussed with reference to the cognitive model of speech production proposed by Valentine et al. [Val96] which predicts a greater difficulty in retrieving names of people.

1. Introduction

Au cours d'une conversation quotidienne, nous produisons en moyenne 100 à 200 mots par minute (Deese [Dee84]). Sachant qu'un adulte normal connaît environ 75000 mots (Oldfiels [Old63]), la récupération, au sein de notre lexique mental[1], des items appropriés constitue donc un phénomène rapide, pour lequel nous nous révélons généralement très efficaces. L'accès lexical n'est toutefois pas infaillible: en particulier, nous nous trouvons quelquefois dans l'impossibilité d'énoncer la forme phonologique d'un mot pourtant bien connu de nous, que nous avons alors l'impression d'avoir «sur le bout de la langue» et à propos duquel nous restons capables de fournir des informations (le nombre des syllabes qui le composent, son phonème initial par exemple). Ce phénomène, appelé «manque du mot», ou «blocage lexical», traduit une perturbation momentanée de l'accès au lexique mental. Tous les types de mot ne sont pas affectés dans les mêmes proportions par les blocages lexicaux, ce qui témoigne d'un degré de difficulté de récupération plus ou moins important. Une catégorie lexicale poserait plus de problèmes que les autres: la catégorie des noms propres[2], au sein de laquelle les noms de personnes seraient davantage touchés.

L'idée selon laquelle l'accès lexical aux noms propres serait relativement malaisé s'appuie sur des données empiriques recueillies auprès de sujets normaux et de certains patients aphasiques[3]. Les premiers doivent généralement remplir un questionnaire (rétrospectif ou bien journalier), dans lequel ils recensent et décrivent leurs blocages lexicaux quotidiens. Selon ces études, les noms propres seraient particulièrement sujets à perturbation puisque 50 à 70 % des manques du mot les concerneraient ! (e.g. Gruneberg et al. [Gru73]; Burke et al. [Bur91, expérience 1]). Quant aux aphasiques, ils souffrent parfois d'un trouble répertorié sous le nom d'»anomie spécifique aux noms propres» (e.g. Semenza et Zettin [Sem89] ; Harris et Kay [Har95] ; Papagno et Capitani [Pap98]). Dans diverses tâches langagières (la production des noms communs notamment), les patients ne rencontrent aucune gêne ; en revanche, lorsqu'ils doivent produire des noms propres, ils se heurtent presque systématiquement à des blocages lexicaux. En l'absence d'un trouble inverse (patients parvenant à retrouver plus facilement les noms propres que les autres unités lexicales), ce type d'anomie étayerait l'idée selon laquelle la récupération des noms propres serait plus « fragile » que la récupération des noms communs[4].

En réalité, si les données rapportées ci-dessus suggèrent bel et bien que les noms propres sont particulièrement difficiles d'accès, elles ne le garantissent nullement. Et ce, pour plusieurs raisons:

En premier lieu, le corpus de résultats expérimentaux relatifs aux sujets «normaux» est encore peu fourni et il est essentiellement constitué, nous l'avons vu, d'études par questionnaires, autrement dit de travaux basés sur des évaluations subjectives et donc peu fiables[5]. Comme le

[1] Le lexique mental ´†est envisagé comme un système dans lequel sont représentées toutes les informations (phonétiques, phonologiques, lexicales, syntaxiques et sémantiques) concernant tous les mots de la langue connue par le sujet et regroupées dans des représentations mentales ou lexicales†ª (Cole [Col89]).

[2] La notion de ´†nom propre†ª renvoie ici exclusivement aux noms propres prototypiques, c'est-à-dire aux noms de personnes (prénoms et noms de famille) et de lieux (noms de pays, de villes, de fleuves, etc.).

[3] Les aphasiques sont des personnes qui, suite ‡ une lésion cérébrale, présentent des troubles langagiers.

[4] Le postulat sous-jacent à une telle supposition est que le langage fonctionne, d'un point de vue cognitif, sur le même modèle chez l'aphasique et le sujet normal.

[5] A notre connaissance, seulement deux études de nature différente ont été effectuées (Rastle et Burke [Ras96]†; Burke et

remarquent Cohen et Faulkner [Coh86], dans ce type d'investigations, «there is no control over many crucial factors» (P. 192). Ainsi, certains blocages, ceux qui sont les plus perturbateurs et/ou les plus mémorables, sont plus susceptibles d'être notés par les sujets que les autres. Par là, la prédominance des manques du mot affectant les noms propres est peut-être simplement liée à un artefact : il est possible que les noms propres retiennent davantage l'attention des personnes testées que les autres items lexicaux car, contrairement à ces derniers, ils ne peuvent pas être remplacés par un synonyme (Cohen et Faulkner [Coh86] ; voir aussi Brédart [Bré93]). Les données récoltées par le biais des questionnaires rétrospectifs sont d'autant moins sujettes à caution que ceux-ci font appel à des capacités mnésiques importantes; les individus doivent en effet retrouver des informations concernant des laps de temps étendus.

En second lieu, chez les aphasiques, la possibilité d'un trouble inverse de l'anomie limitée aux noms propres se laisse entrevoir. En effet, certains patients semblent éprouver des difficultés nettement moindres dans la récupération des noms propres que des autres types de mots (e.g. Semenza et Sgaramella [Sem93]). Ces cas de préservation sélective, en général encore trop succinctement décrits, invitent à substituer à la notion de difficulté d'accès aux noms propres celle de spécificité d'accès.

En conclusion, il convient de mettre en place des investigations supplémentaires afin de confirmer ou d'infirmer l'hypothèse de difficulté associée à la récupération des noms propres. L'étude rapportée ici, effectuée auprès de sujets normaux, vise un tel objectif. Il s'agit d'une expérience informatisée de dénomination[6] à partir de photographies (de célébrités et d'objets usuels). Nous avons comparé l'accessibilité aux noms propres et aux noms communs par une mesure des temps de réponse et du nombre des blocages lexicaux.

2. Méthode[7]

2.1 Sujets

Les sujets, au nombre de 98, comprennent 55 femmes et 43 hommes « normaux », droitiers, âgés de 19 à 75 ans (et répartis de manière équilibrée dans trois groupes d'âges: 35 «sujets jeunes» ont entre 19 et 34 ans; 30 «sujets d'âge moyen» ont entre 36 et 54 ans; enfin, 33 «sujets âgés» ont entre 55 et 75 ans). Ils ont tous un niveau de scolarité égal ou supérieur au baccalauréat. Leur vision est correcte, avec ou sans correction (lunettes, verres de contact). Enfin, ils sont de langue maternelle française et ont passé la plus grande partie de leur vie en France; cette précaution s'imposait, l'expérience impliquant des connaissances linguistiques et culturelles.

2.2 Stimuli

Les stimuli-cible comprennent:
- 16 photographies de visages célèbres (e.g. Johnny Hallyday; Catherine Deneuve; Claire Chazal) ;
- 16 photographies d'objets de la vie courante (e.g. un piano; une carafe; un manteau).

Lors de la sélection des cibles, nous avons pris en considération un certain nombre de variables linguistiques et psycholinguistiques susceptibles d'avoir un impact sur le processus de dénomination et donc de biaiser les résultats. En ce qui concerne les items lexicaux, nous avons veillé à ne choisir que des noms d'objets et de célébrités très familiers (voir Evrard et al. [Evr99]) et nous avons apparié au mieux les noms communs avec les noms propres quant au nombre de syllabes et de phonèmes[8]. Pour ce qui est des images, elles sont en couleur. Les visages et les objets y apparaissent tous sans contexte, sur un même fond bleu.

2.3 Matériel

Nous avons utilisé un ordinateur portable Macintosh avec écran couleur. Un microphone avec casque de support était branché sur l'appareil. Un logiciel a été spécialement conçu pour l'étude. L'appareillage permet l'affichage des stimuli (croix de fixation, images), le contrôle de leur durée de présentation, la détermination et l'enregistrement des temps de réaction (détection des réponses vocales).

2.4 Consigne et déroulement

Le sujet, interrogé individuellement, avait pour consigne de nommer oralement aussi rapidement que possible les objets (production de noms communs) et les célébrités (production de noms de famille) auxquels il serait confronté. Nous soulignions la nécessité de ne faire aucun commentaire durant tout le déroulement de l'épreuve, même (voire surtout!) en cas de «mot sur le bout de la langue», et de bannir tout bruit parasite.

Les images apparaissaient une à une sur l'écran, dans un ordre aléatoire et variable d'un sujet à l'autre.

---

al. [Bur91, expérience 2]). Elles reposent l'une et l'autre sur une tâche de dénomination à partir de définitions et de descriptions mais aboutissent à des résultats contradictoires: la première est favorable aux conclusions des études par questionnaires, tandis que la seconde les discrédite.

[6] La dénomination peut être définie comme la désignation d'un être ou d'une chose extralinguistique par un nom.

[7] Les données rapportées dans cet article ont été recueillies au cours d'une expérience de plus grande ampleur centrée sur les effets de l'amorçage sémantique et phonologique sur la dénomination. Seuls les résultats relatifs aux images non amorcées sont pris en considération ici.

[8] Le contrôle de la fréquence d'utilisation n'a pu être qu'intuitif Étant donné qu'il n'existe pas de tables de fréquence pour les noms propres.

Une photographie donnée ne disparaissait que lorsque l'ordinateur détectait un son ou sur commande de l'expérimentateur (en cas d'absence de réponse prolongée[9], il suffisait d'appuyer sur une touche quelconque du clavier afin d'effacer l'image).

L'ordinateur mesurait et enregistrait automatiquement les temps de réponse (correspondant aux laps de temps séparant le début de la prononciation de l'apparition d'une image donnée) et nous notions, au fur et à mesure de leur apparition, les réponses inadéquates (les mots que nous n'attendions pas, les absences de réponse, etc.).

Enfin, pour identifier les blocages lexicaux, après le test à proprement parler, nous soumettions au sujet chacune des photographies qu'il n'avait pas nommées. Nous lui demandions:
1) S'il reconnaissait la célébrité / l'objet représenté(e);
2) Le cas échéant, s'il était capable de fournir des renseignements à son sujet (e.g. sa profession pour les célébrités);
3) S'il pouvait donner son prénom (pour les célébrités) et/ou des informations phonologiques relatives à son nom (e.g. lettre initiale, nombre de syllabes).
Lorsque les réponses aux deux dernières questions étaient satisfaisantes, nous considérions que le sujet s'était heurté à un blocage lexical lors du test proprement dit.

3. Résultats

Les données récoltées se répartissent ainsi: 86,02 % constituent des bonnes réponses, 7,35 % des manques du mot, 4,75 % des «erreurs» (e.g. absence de reconnaissance de l'objet ou de la célébrité, production d'un mot non attendu). Enfin, 1,88 % renvoient à un problème de détection sonore engendré par exemple par un bruit parasite.

Les résultats rapportés ci-dessous concernent exclusivement les réponses correctes, pour lesquelles nous avons étudié les temps de réaction, et les manques du mot selon l'appartenance lexicale du mot-cible (nom propre versus nom commun) au moyen d'analyses de variance.

3.1 Latences de réponse

Nous avons éliminé les temps de réaction aberrants de chaque sujet, c'est-à-dire inférieurs ou supérieurs au temps de réaction moyen + ou - deux écarts types. Nous avons alors pu effectuer une nouvelle estimation du temps de réaction moyen par sujet. Celui-ci atteint presque 2 secondes (1982 ms) pour le nom propre alors qu'il est inférieur à une seconde et demi pour le nom commun(1318 ms). Cette différence est largement significative: $F(1,98) = 198{,}476$, $p<0{,}0001$.

3.2 Blocages lexicaux

Les manques du mot sont significativement plus fréquents lorsque la cible est un nom propre que quand il s'agit d'un nom commun ($F(1,98) = 88{,}928$, $p<0{,}0001$). Ainsi, les sujets ont en moyenne 13,4 manques du mot pour 100 cibles noms propres et seulement 1,3 pour 100 cibles noms communs!

4. Discussion

Les résultats recueillis, qu'ils se rapportent aux réponses correctes ou bien aux blocages lexicaux, vont dans le sens des conclusions tirées à partir des études par questionnaires. D'une part, le commencement plus tardif de la prononciation du nom de personne que du noms commun peut refléter un processus cognitif d'accès lexical au nom propre plus long, et par là un degré d'accessibilité moins élevé. D'autre part, la plus forte proportion des manques du mot affectant les noms de personnes tend à prouver que ces items engendrent bel et bien davantage de problèmes de récupération.

Il convient néanmoins de souligner le caractère moins fiable des temps de réaction que des manques du mot. En effet, les premiers, contrairement aux seconds, ne concernent pas spécifiquement l'accès au lexique mental. Ainsi, durant le laps de temps qui sépare l'apparition de la photographie de la proposition de réponse, le sujet effectue diverses opérations cognitives, parmi lesquelles la recherche dans le lexique interne ne constitue qu'une étape; il doit, par exemple, procéder à un traitement visuel de l'image, reconnaître l'item représenté, etc. Ainsi, une latence moyenne plus étendue pour les noms propres ne reflète pas indubitablement une durée plus importante de l'accès lexical en lui-même. Elle peut par exemple s'expliquer, au moins partiellement, par un temps d'analyse visuelle plus long, et ce, d'autant plus que les visages constituent des exemplaires relativement semblables, sans doute plus difficiles à différencier les uns les autres et à identifier que les objets. En résumé, les temps de réaction sont pour nous plus secondaires et moins parlants parce qu'ils ne sont pas exclusivement liés à la récupération lexicale proprement dite. Quoiqu'il en soit, ceux que nous avons récoltés sont en harmonie avec l'interprétation basée sur les manques du mot; par là, ils la renforcent.

Nous pouvons nous demander pourquoi nos résultats ne s'accordent pas à ceux de Burke et al. [Bur91, expérience 2] qui ne recensent pas davantage de blocages pour les noms propres que pour les noms communs (voir ci-dessus). A notre avis, deux grandes différences séparent leur étude de la nôtre. La première concerne le type de tâche mis en œuvre: pour provoquer la dénomination, ces auteurs s'appuient sur des définitions et des descriptions verbales, alors que nous utilisons des photographies. La seconde distinction a trait au type de stimuli verbaux employé: par exemple, le protocole de Burke et al. [Bur91] comprend des mots-cibles rares, le nôtre des mots-cibles familiers.

[9] C'est-à-dire égale ou supérieure à 10 secondes, durée qui nous a paru raisonnable.

L'écart des résultats entre les deux études ne s'explique sans doute pas par la nature de la tâche: Rastle et Burke [Ras96] se basent, tout comme Burke et al. [Bur91], sur une expérience de dénomination à partir de définitions mais aboutissent aux mêmes conclusions que nous. La différence des résultats provient donc certainement des stimuli-cibles eux-mêmes (ceux-ci sont partiellement différents dans les études de Burke et al. [Bur91] et de Rastle et Burke [Ras96]). Des expériences complémentaires seraient utiles pour vérifier une telle hypothèse et mieux cerner les caractéristiques des mots-cibles entrant en jeu dans les variations de l'accessibilité aux deux catégories lexicales. Notons au passage que des études de laboratoire supplémentaires seraient également nécessaires afin de contrôler si les noms de personnes sont, comme le laissent supposer les études par questionnaires, les plus difficiles à récupérer des noms propres.

Pour finir, envisageons les mécanismes cognitifs qui pourraient différencier la récupération du nom propre de la récupération du nom commun. Nous présenterons ci-dessous succinctement le modèle psycholinguistique de Valentine et al. [Val96], qui prévoit une plus grande vulnérabilité de l'accès lexical aux noms de personnes et fournit donc un cadre explicatif à nos résultats. Valentine et al. [Val96] comparent les opérations cognitives permettant, d'une part la dénomination de l'objet, et d'une part la dénomination du visage :

Dans le cas d'un objet (e.g. une voiture) présenté visuellement, une unité de reconnaissance est activée. Elle donne accès à la mémoire sémantique, et plus précisément aux traits sémantiques qui constituent le sens du mot à retrouver (e.g. «permet de se déplacer», «a quatre roues», etc.). C'est à partir de ces nombreuses informations sémantiques que l'item lexical cible peut être sélectionné. Il a d'abord une forme abstraite, pré-phonologique : le «lemma». La forme phonologique du mot cible, appelée «lexème», est ensuite récupérée. Un programme articulatoire rend alors possible la prononciation du nom commun. La présentation d'un visage conduit également à l'activation d'une unité de reconnaissance. Comme les noms de personnes ont un référent unique et ne renvoient pas, contrairement aux noms communs, à des concepts, l'unité de reconnaissance du visage active, non plus le système sémantique, mais un «nœud d'identité de la personne» ou «token marker». L'accès au lemma, puis au lexème se produit à partir de cette unité unique.

Ainsi, le lemma d'un nom commun est connecté à de nombreuses unités sémantiques qui interviennent dans sa récupération. Au contraire, le lemma d'un nom de personne n'est activé que par le biais d'une seule unité (le «nœud d'identité de la personne»), de sorte que son accès est particulièrement fragile. Reste à élaborer des expérimentations visant à évaluer la validité d'une telle interprétation...

Je tiens à remercier particulièrement Marc Lafon qui a construit le logiciel utilisé dans cette étude.

Bibliographie

[Bré93] Brédart S. (1993), « La production des noms propres », Revue de Neuropsychologie, Vol. 3(2), pp. 203-220.

[Bur90] Burke D.M. et Laver G.D. (1990), «Aging and word retrieval : Selective age deficits in language». In E.A. Lovelace [ED.], Aging and Cognition : Mental Processes, Self Awareness and Interventions. North Holland : Elsevier, pp. 281-300.

[Coh86] Cohen G. et Faulkner D. (1986), «Memory for proper names : Age differences in retrieval», British Journal of Developmental Psychology, Vol. 4, pp. 187-197.

[Cole89] Cole P. (1989). « Morphologie et accés au lexique », Lexique, Vol. 8, pp. 65-85.

[Dee84] Deese, J. (1984), Thought into speech: The psychology of language, Englewood Cliffs (NJ), Prentice-Hall.

[Evr99] Evrard M., Ferrati F. et Nespoulous J.L. (1999), «La familiarité dans l'accès lexical: le cas des noms propres». Poster présenté à Soulac (Gironde), au colloque «Jeunes Chercheurs en Sciences Cognitives», pp. 258-259.

[Gru73] Gruneberg M.M., Smith R.L. et Winfrow P. (1973), «An investigation into response blockaging», Acta Psychologica, Vol. 37, pp. 187-196.

[Har95] Harris D.M. et Kay, J. (1995), « Selective impairment of the retrieval of people's names : A case of category specificity », Cortex, Vol. 31, 575-582.

[Old63] Oldfield R.C. (1963), «Individual vocabulary and semantic currency: A preliminary study», British Journal of Social and Clinical Psychology, Vol. 2, pp. 122-130.

[Pap98] Papagno C. et Capitani E. (1998), «Proper name anomia: A case with sparing of the first-letter knowledge», Neuropsychologia, Vol. 36(7), pp. 669-679.

[Ras96] Rastle K.G. et Burke D.M. (1996), « Priming the Tip of the Tongue : Effects of Prior Processing on Word Retrieval in Young and Older Adults », Journal of Memory and Language, Vol. 35, pp. 586-605.

[Sem93] Semenza C. et Sgaramella T.M. (1993), «Production of proper names : A clinical case study of the effect of phonemic cueing», Memory, Vol. 1(4), pp. 265-280.

[Sem89] Semenza C. et Zettin M. (1989), «Evidence from aphasia for the role of proper names as pure referring expressions», Nature, Vol. 342, pp. 678-679.

[Val96] Valentine T., Brennen T. et Brédart S. (1996), The cognitive psychology of proper names, London : Routledge.

Ecriture et dyslexie : approche phonologique

Carine Sabater, Virginie Daffaure, Sonia De Martino et Véronique Rey

Université de Provence, Laboratoire Parole et Langage, ESA 6057 CNRS
29, Avenue Robert Schuman, 13621 Aix-en-Provence

ABSTRACT

The main objective of this study is to know if mistakes made by 19 phonological dyslexic children in a 30 no-words dictation recover from a linguistic analysis or no. The results obtain in articulation allow us to assert that these subjects have an implicit phonological awareness Moreover, we have tried to realize a quantitative and a qualitative analysis of mistakes made in the dictation. The results of the metalinguistic exercices and of the dictation suggest a deficit of an explicit phonological awareness.

1. INTRODUCTION

Le déficit en conscience phonologique est une des caractéristiques de la dyslexie développementale. Certains auteurs ont proposé l'hypothèse d'un déficit temporel non spécifique comme cause au déficit phonologique [Tal 73][Tal 80]. D'autres auteurs ont avancé que ce déficit phonologique est essentiellement de nature linguistique [Mod 97]. Afin de mieux comprendre le type d'erreurs des enfants dyslexiques, nous nous sommes fixées comme objectif d'expliquer les troubles de la conscience phonologique ainsi que leurs conséquences sur la production d'écrit à partir d'une évaluation faite auprès de 19 enfants dyslexiques. Nous avons tenté d'abord de voir sur quels exercices de conscience phonologique les sujets éprouvent le plus de difficultés puis, nous avons examiné leurs erreurs en production écrite à partir d'une grille d'analyse et enfin, avons décidé si ces fautes relèvent d'une étude linguistique ou non.

2. METHODE

L'expérience que nous avons menée auprès de ces enfants dyslexiques a consisté en une évaluation de leur conscience phonologique et en une analyse linguistique de leurs fautes dans les conversions phonèmes-graphèmes.

2.1 Les Sujets

Les 19 enfants que nous avons testés (12 garçons et 7 filles) étaient âgés de 8 à 11 ans. Chacun d'entre eux a été sélectionné par sa propre orthophoniste, qui attestait que l'enfant ne souffrait pas de problèmes sensoriels primaires ni de troubles psychologiques graves.

2.2 Les Tâches

La première partie de l'expérience est un examen de l'articulation (qui est utilisé par les phoniatres en cabinet) et qui consiste à répéter des syllabes à structure CV, des syllabes ou des non-mots composés de diconsonnes (clusters), des semi-voyelles, des voyelles et leurs correspondantes nasales, des mots et des phrases.

La deuxième partie est une évaluation de la conscience phonologique testée au moyen d'exercices métaphonologiques portant sur des mots ou des logatomes :

- trouver parmi trois mots les deux qui ont le même phonème consonantique au début ou au milieu,
- trouver parmi trois mots les deux qui riment,
- trouver parmi trois non-mots l'intrus qui ne contient pas le même phonème consonantique au début (opposition sur le voisement et le lieu d'articulation),
- trouver parmi trois non-mots l'intrus qui ne contient pas le même phonème consonantique au milieu (opposition sur le voisement et le lieu d'articulation),
- compter le nombre de fois qu'un phonème donné est entendu dans un mot,
- compter le nombre de phonèmes contenus dans un non-mot donné.

Enfin, la troisième partie est une dictée de 30 logatomes à structure syllabique simple ou complexe. Cette activité vise à tester les capacités analytiques de l'enfant à transcrire un non-mot. Il est contraint ici de recourir à la transcription phonème-graphème.

3. RESULTATS

3.1 Bilan articulatoire

Afin de déterminer un éventuel trouble du langage oral, nous avons fait passer aux enfants un bilan articulatoire. Les scores qui varient de 72 à 84/87 ne reflètent pas de graves troubles du langage oral. Nous pouvons noter tout de même de mauvais résultats en répétition d'associations consonantiques (ils font beaucoup d'inversions phonémiques, ex : spa est répété psa). Nous avons également observé de nombreuses confusions sur le voisement des phonèmes (assourdissement et sonorisation).

3.2 Conscience phonologique

Les scores, qui varient de 20 à 51/62, témoignent d'une grande hétérogénéité au niveau métaphonologique chez ces enfants. Toutefois, nous avons pu nous apercevoir que les difficultés variaient selon plusieurs critères : selon la nature du mot d'une part (mot réel ou non-mot) et selon la place du phonème sur lequel on travaille (début, milieu, fin), d'autre part.

Nous avons donc pu constater que :
- les performances sont meilleures sur les mots que sur les non-mots.
- les performances sont meilleures si le phonème est au début plutôt qu'au milieu du mot.
- en ce qui concerne les logatomes, le voisement et le lieu sont mieux repérés au début qu'au milieu du mot.
- pour l'exercice où il faut compter le nombre de fois qu'un phonème est entendu dans un mot, les sujets se heurtent au problème où deux phonèmes ne diffèrent que par le voisement. Par exemple, nombreux sont ceux qui ont affirmé qu'ils entendaient deux fois le son /t/ dans le mot « digital ».

3.3 Les erreurs en production écrite

Le principal objectif de notre étude était de voir si les fautes produites lors de la dictée de logatomes étaient de nature linguistique ou non.
Pour répondre à cette question, nous nous sommes proposée de reprendre la classification des fautes que Sheila E. Blumstein [Blu 95] a dressée lorsqu'elle a travaillé sur les erreurs de production orale d'enfants souffrant d'aphasie développementale. Elle compte quatre catégories d'erreurs :
1) Erreurs de substitutions de phonèmes.
2) Erreurs de simplification où un phonème ou une syllabe est supprimée.
3) Erreurs d'addition où un phonème extérieur ou une syllabe est ajoutée au mot.
4) Erreurs d'environnement où l'occurrence d'un phonème particulier est influencée par le contexte phonétique environnant :
* phénomènes de métathèses.
* erreurs d'assimilation progressive et régressive.

A la suite de l'examen de chaque dictée des enfants, nous avons pu dresser le tableau suivant dans lequel nous faisons apparaître le pourcentage de fautes relatif à chaque catégorie d'erreurs selon S. E. Blumstein. Ainsi, la première et la quatrième catégories seront regroupées sous le terme de fautes phonologiques, la deuxième sous le terme d'omissions et la troisième concernera les insertions.

Table 1 : Pourcentages d'erreurs en dictée.

Fautes phonologiques	Omissions	Insertions
48%	39%	13%

Il est aisé de constater que le pourcentage de fautes phonologiques est largement supérieur à celui des deux autres types d'erreurs, si l'on regarde les résultats dans leur globalité et non au cas par cas. Mais ces chiffres semblent bien refléter la tendance générale puisque 15 sujets sur 19 font une majorité de fautes phonologiques parmi lesquels 3 ne commettent que ce type d'erreurs. Les cinq autres enfants font une majorité d'omissions. Aucun enfant ne commet plus d'insertions que de fautes phonologiques ou d'omissions.

En ce qui concerne les fautes phonologiques, nous avons pu noter que la majorité d'entre elles se produisait au milieu du mot. Ceci concerne 17 enfants et corrobore donc ce que nous avions dit pour la conscience phonologique. Nous nous attarderons plus longuement sur ce type de fautes lors du deuxième classement.
Pour les omissions, le plus grand nombre d'erreurs se retrouve encore au milieu du mot, mais 4 enfants ne commettent aucun oubli. Les omissions ou ablations peuvent toucher le début du mot (c'est une aphérèse et elles représentent ici 32%), le milieu du mot (58%), ou la fin du mot (c'est une apocope, 10%).
Nous retrouvons toujours la même tendance pour les insertions qui se produisent en majorité au milieu du mot. Comme les omissions elles peuvent toucher n'importe quelle place du mot : si elle se fait à l'avant, c'est une prothèse (1.15% des cas ici), à l'intérieur du mot, c'est une épenthèse (81.60%) mais elle peut également se faire à la fin du mot (17.24%).
A la suite de cette analyse des différents types d'erreurs relevés dans la dictée de logatomes, nous pouvons donc affirmer que les fautes phonologiques sont nettement plus nombreuses que les omissions ou les insertions. De façon générale, ces trois types d'erreurs apparaissent majoritairement au milieu du mot. Nous avons pu remarquer que les enfants "oublient" ou insèrent des phonèmes dans le seul but de rendre la transcription plus facile, c'est-à-dire de faire en sorte que la structure syllabique du logatome ressemble le plus possible à celle d'un mot réel. En effet, les omissions touchent essentiellement les phonèmes consonantiques qui appartiennent à une diconsonne. Par contre, les sujets intègrent majoritairement des voyelles dans le groupe consonantique afin de retrouver une syllabe à structure CVC.
Cependant, même s'il paraît vraisemblable que ces trois formes d'erreurs sont commises, explicitement ou non, à faciliter la transcription phonème-graphème, nous avons également émis l'hypothèse que l'entourage phonologique du mot pouvait influencer ces fautes. En effet, en analysant les productions écrites, nous avons été confrontées à des phénomènes d'anticipation et de persévération. Les anticipations consistent à déplacer un trait d'un phonème de la fin du mot sur un phonème du début. Les anticipations ne provoquent que des fautes phonologiques, que ce soit des phénomènes de voisement (les phonèmes non voisés de la fin du mot influencent leurs pendants voisés ; ainsi il est fréquent de constater que /g/ devient /k/ dans « golqué »et /d / devient /t / dans « drastron ») ou de métathèse où il y a déplacement d'un phonème : ainsi « fulger » devient « fluger ». Les persévérations, par contre, consistent à déplacer un trait d'un phonème du début du mot sur un phonème de la fin. Elles peuvent provoquer des fautes phonologiques, des omissions et des insertions. Par exemple, un phonème en fin de mot est remplacé par un autre qui apparaît avant lui, comme c'est le cas dans le logatome « flachu » qui est transcris « phlafu ». Ainsi, les phénomènes de

persévération et d'anticipation sont très présents dans les productions des sujets.

Notre but second est de voir sur quel type de trait phonologique portent les erreurs en production écrite. Pour cela, nous avons établi une hiérarchie de traits :
* le lieu d'articulation comprenant les traits suivants : bilabial, labiodental, dental, alvéolaire, post-alvéolaire, palatal, vélaire et uvulaire.
* le mode d'articulation comprenant les traits suivants : occlusif, nasal, constrictif, approximant et latéral.
* le voisement des phonèmes : ils peuvent être sonores (voisés), c'est-à-dire qu'il y a vibration des cordes vocales, ou sourds (non voisés), c'est-à-dire qu'il y a absence de telles vibrations.
* la classe des métathèses regroupera les erreurs portant sur les déplacements des phonèmes : interversion de deux phonèmes ou déplacement d'un seul.
Pour cette seconde classification, nous n'avons pris en compte que les erreurs relatives aux phonèmes consonantiques. Après un examen quantitatif de ces fautes, nous avons pu établir le classement suivant qui répertorie, dans un ordre décroissant, le pourcentage de fautes relatif à chaque trait phonologique :
1) Erreurs de voisement : 36.12%
2) Erreurs de lieu : 30.19%
3) Erreurs de mode : 18.87%
4) Métathèses : 14.82%

Les erreurs de voisement sont celles que nous avons rencontrées en majorité. Sur la totalité des sujets, nous avons relevé 56% de sonorisations et 44% d'assourdissements. L' erreur la plus fréquente est la transformation du /k/ en /g/ et inversement. Comme nous le verrons ultérieurement à propos des erreurs de lieu, ce sont les phonèmes vélaires qui sont les plus propices à être modifiés, et la tendance générale est le voisement des phonèmes sourds. Le nombre élevé d'erreurs de voisement conforte l'analyse des orthophonistes car elles reconnaissent que les enfants dyslexiques éprouvent de grosses difficultés à percevoir si un phonème est sourd ou sonore. Nous pouvons même dire que c'est l'un des principaux symptômes, au niveau linguistique évidemment, de la dyslexie.
Les erreurs de lieu peuvent concerner un déplacement de celui-ci vers l'arrière (38% des cas) ou vers l'avant du conduit vocal (62% des cas). Les consonnes les plus touchées par un déplacement du lieu d'articulation vers l'arrière sont en premier lieu les dentales, puis les bilabiales et les labiodentales. Celles qui ont la plus grande chance d'avancer sont les vélaires, les alvéolaires et les post-alvéolaires. Ainsi, nous pouvons noter que les phonèmes qui posent le plus de difficultés aux enfants sont sans conteste ceux qui sont le plus à l'arrière du conduit vocal (et en particulier les vélaires), puisque la grande majorité des sujets a tendance à avancer le lieu d'articulation.
Les erreurs de mode sont, en nombre, minoritaires du point de vue des traits distinctifs. Aucun enfant ne commet en majorité ce type de fautes. Dans un ordre décroissant, les phonèmes les plus touchés sont les occlusifs, puis les constrictifs et enfin les latéraux. Les consonnes occlusives et les latérales ont tendance à devenir constrictives et les constrictives à devenir occlusives. Ce sont donc les phonèmes occlusifs les plus enclins à être modifiés, sans doute à cause de leur nature intrinsèque.

Tout au long de cette analyse en termes de traits de lieu, de mode et de voisement, nous avons déclaré quel était le pourcentage de chaque type de fautes ne concernant qu'un seul trait distinctif. Nous nous proposons maintenant de donner brièvement, et à titre indicatif, le pourcentage de types de fautes touchées par deux ou par les trois traits phonologiques.

Table 2 : Pourcentages d'erreurs relatifs à un, deux ou trois traits distinctifs.

Erreurs d'un trait	Erreurs de deux traits	Erreurs de trois traits
13% concernent le lieu. 1% concerne le mode. 46% concernent le voisement.	23% concernent le lieu et le mode. 8% concernent le lieu et le voisement.	9% concernent le lieu, le mode et le voisement.

Les fautes les plus nombreuses sont donc celles qui portent sur un seul trait distinctif et particulièrement celui du voisement.
Ainsi que nous l'avons dit précédemment, ces sujets nous ont prouvé, au cours du bilan articulatoire, qu'ils n'éprouvaient pas de difficultés spécifiques à répéter des mots, des non-mots et des phrases. Ceci démontrerait donc qu'ils possèdent un système phonologique implicite de la langue. Pourtant, lors du passage à la traduction graphémique qui est nécessairement explicite, ils "oublient" majoritairement un seul trait phonologique. Il semblerait que la conversion phonème-graphème oblige une réduction de la variation des phonèmes : le graphème "b" ne traduit que le phonème /b/ dans les logatomes. La catégorie phonologique se trouverait donc renforcée par le passage à l'écriture, mais ceci ne semble pas être le cas chez les enfants dyslexiques.

Les métathèses sont la catégorie de fautes minoritaires du point de vue quantitatif. Seul un enfant commet ce type d'erreurs de manière majoritaire. Les métathèses peuvent se traduire par une interversion entre deux phonèmes consonantiques contigus ou éloignés, ou entre une consonne et une voyelle, ou même être le déplacement d'un seul phonème. Nous n'avons cependant pas relevé d'interversion entre deux phonèmes consonantiques éloignés. Ceci prouve donc que toutes les métathèses touchant les consonnes se produisent exclusivement sur des diconsonnes (clusters) dans le but de rendre la transcription plus facile.

Les résultats que nous venons de présenter montrent que le déficit phonologique se manifeste non seulement dans des exercices de conscience phonologique mais aussi dans l'encodage des graphèmes et dans la capacité à juger de l'ordre temporel (manifesté par les fautes dans les métathèses). Ainsi, un enfant peut très bien donner une production spontanée sans erreurs phonologiques mais révéler un déficit phonologique grave dans son enregistrement des phonèmes.

CONCLUSION

Ainsi que nous l'avons vu tout au long de cette analyse, cette difficulté à transcrire les phonèmes dans un ordre précis serait peut-être lié à un déficit du jugement de l'ordre temporel puisque nous avons pu noter de nombreuses métathèses. Mais il semblerait également que l'influence phonologique produite à l'oral (en situation de coarticulation de parole spontanée) est, chez eux, transcrite à l'écrit alors que ces catégories graphémiques correspondantes devraient être "étanches". En effet, nous avons pu être confrontées à de nombreux phénomènes d'assimilation où un trait distinctif d'un phonème influence son homologue sur un autre phonème. Par exemple, le non-mot "paldo" pourra être répété correctement mais lors de la transcription, il apparaît sous la forme de "palto". Nous avons émis l'hypothèse que le phonème /p/ qui est non voisé a influencé le phonème /d/ qui est voisé et qui est remplacé par son pendant non voisé. Par conséquent, même si le système phonologique à l'oral paraît relativement stable, il n'en est pas de même lors de la conversion phonème-graphème. Il semblerait que pour les enfants dyslexiques, les graphèmes s'influencent à l'écrit comme les phonèmes s'influencent à l'oral.

La grille d'analyse élaborée par Sheila E. Blumstein paraît tout à fait appropriée pour l'examen des fautes en production écrite de sujets dyslexiques bien qu'elle ait été établie au départ pour une pathologie du langage oral. Cela prouve que les fautes relevées dans l'enregistrement écrit des enfants dyslexiques présentent des caractéristiques existant à d'autres niveaux dans d'autres pathologies du langage. Ainsi, il s'agit bien en premier lieu d'un déficit de nature linguistique. Ceci n'exclut pas le fait que ce déficit linguistique pourrait être en interaction avec un déficit du traitement temporel.

BIBLIOGRAPHIE

[Ale 89] Alegria J., Morais J., (1989), "Analyse segmentale et acquisition de la lecture", dans Rieben L; et Perfetti C., L'apprenti lecteur : recherches empiriques et implications pédagogiques. Neuchâtel : Delachaux et Niestlé, pp. 173-192.

[Ale 96] Alegria J., Morais J., (1996), "Métaphonologie, acquisition du langage écrit et troubles associés" dans Carbonnel S., Approche cognitive des troubles de la lecture et de l'écriture chez l'enfant et l'adulte, Marseille, Solal Editeurs, pp.81-93.

[Blu 95] Blumstein S. E., (1995), "The Neurobiology of the sound structure of language", Language, The Cognitive Neurosciences, M. S. GAZZANIGA (Ed), Massachusetts Institut of Technology, pp. 915-929.

[Bra 91] Bradley L., Bryant P.E, (1991), "Phonological skills before and after learning to read", dans Brady S.A. & Shankweiler D.P., Phonological processes in literacy. Hillsdale, N.J : LEA, pp.37-46.

[Ehr 89] Ehri L.C, (1989), "Apprendre à lire et à écrire les mots", dans Rieben L. et Perfetti C., L'apprenti lecteur : recherches empiriques et implications pédagogiques. Neuchâtel : Delachaux et Niestlé, pp.103-128.

[Mod 97] Mody M., Studdert-Kennedy M;, Brady S., (1997), "Speech perception deficits in poor readers : auditory processing or phonological coding ?", Journal of Experimental Child Psychology, 64, pp.199-231.

[Mor 87] Morais J., Alegria J., Content A., (1987), "The relationship between segmental analysis and alphabetic literacy : an interactive view". European Bulletin of Cognitive Psychology, 7, pp.415-438.

[Tal 73] Tallal P., Piercy M., (1973), "Defects of non verbal auditory processing in children with developmental aphasia", Nature, 241, pp. 468-469.

[Tal 80] Tallal P; (1980), "Auditory temporal perception,phonics, and reading disabilities in children", Brain & Language, 9, pp. 182-198.

[Tun 89] Tunmer W.E (1989), "Conscience phonologique et acquisition de la langue écrite" dans Rieben L; et Perfetti C., L'apprenti lecteur : recherches empiriques et implications pédagogiques. Neuchâtel : Delachaux et Niestlé, pp.197-215.

[Val 96] Valdois S., "Les dyslexies développementales", dans Carbonnel S., Approche cognitive des troubles de la lecture et de l'écriture chez l'enfant et l'adulte, Marseille, Solal Editeurs, pp.137-149.

Equivalence motrice et dominance hémisphérique
Le cas de la voyelle [u]. Étude IRMf

Monica BACIU, Christian ABRY°, Christoph SEGEBARTH#*

*LPE UMR CNRS 5105/°ICP UMR CNRS 5009/#CHU-Grenoble INSERM 438
LPE Université Pierre Mendès-France BP 47 F-38040 Grenoble Cédex 9
Tél.: ++33 (0)4.76.82.78.07- Fax: ++33 (0)4.76.82.78.34
Monica.Baciu@upmf-grenoble.fr/abry@icp.inpg.fr

ABSTRACT

Speech motor equivalence, with sound (formant) equifinality, has been demonstrated for the universal vowel [u]. 7 right handed normal subjects were trained to perform compensatory articulation for [u], the acoustically relevant constriction at their lips being impeded by a lip tube. 3 tasks were under fMRI examination : (i) normal articulation, (ii) compensatory articulation and (iii) normal/compensatory perceptual expectation. The overall difference between normal and compensatory articulation was a clear change in hemispheric dominance : left perisylvian language activation was shifted to right activation in the homologues for production (inferior frontal gyrus, say Broca) and perception (temporal gyri). The *perceptual* task evidenced both a right and left activation of *production* area (inferior frontal gyrus, Broca again). These right hemispheric activations could correspond to a subcategorization process of linguistic sounds : compensatory motor equivalent [u] being a phonetic subcategory of the phonological [u] sound prototype.

1. INTRODUCTION

Le but de cette étude est d'examiner l'activité cérébrale dans une tâche d'équivalence motrice pour la production d'une voyelle. L'équifinalité est l'un des paradigmes les plus révélateurs pour comprendre la plasticité des actes moteurs en général. Pour la parole, en particulier, il est dans sa nature même d'offrir des variantes en fonction des exigences de la coarticulation. La méthode des perturbations a été appliquée aux sons de la parole de manière statique ou dynamique. Pour les voyelles, des perturbations ont déjà été effectuées sur un effecteur proximal comme la mandibule. Mais pas sur un effecteur final, soit l'articulateur qui produit la constriction cruciale, typiquement la langue contre le palais dur pour [i]. Pour [u], par contre, il a été démontré qu'en perturbant la constriction principale, aux lèvres, soit en l'ouvrant par un tube, et à condition qu'en compensation la constriction linguale soit fortement reculée, dans ce cas les valeurs des deux premiers formants F1-F2, caractéristiques de cette voyelle, pouvaient être atteintes ([Sav95] et [Sav99] contiennent les références des expériences antérieures). Même si ce [u] compensé n'est pas perceptivement un succès total en équivalence acoustique (il ne l'est que si l'on modifie la relation avec F1 de la fréquence laryngienne Fo), c'est à cette tâche — essayer d'approcher au mieux cette forme acoustique — que nous avons pu entraîner des sujets non naïfs. Nous espérons tirer de cette étude des enseignements non seulement sur l'*équifinalité* en parole, mais sur les processus de *catégorisation* en œuvre pour passer des sons non déjà catégorisés prototypiquement dans la phonologie de la langue — qui sont donc des exemplaires ou des sous-catégories — aux catégories encodées phonologiquement.

2. MATERIEL ET METHODES

Nous avons examiné 7 sujets sains droitiers, 2 femmes et 5 hommes. Tous étaient des sujets entraînés.

Chaque sujet a subi un seul examen IRMf. Un paradigme de type *block* a été appliqué pendant trois scans fonctionnels. Chaque scan a consisté en l'alternance d'" époques " de contrôle et d'" époques " de tâche. L'alternance contrôle-tâche a été répétée quatre fois. Le passage d'une époque de contrôle à une époque de tâche était signalé par une consigne visuelle écrite en capitales, que le sujet allongé dans l'aimant recevait depuis un ordinateur relié à un projecteur via un système de miroirs.

2.1. Paradigme

Pendant le premier scan fonctionnel nous avons testé la prononciation normale du [u]. Pendant les époques de contrôle, les sujets étaient au repos et fixaient un point au centre de l'écran. Pendant les époques de tâche, les sujets articulaient leur [u] selon leur prononciation normale. Cette prononciation était réalisée avec une articulation statique ; mais sans phonation, de façon à ne pas induire des mouvements (plus précisément glotte ouverte, pour que les sujets puisse respirer doucement pendant la tenue de la voyelle).

Pendant le deuxième scan fonctionnel nous avons testé la prononciation compensée du [u], prononciation effectuée pendant que les sujets avaient un court tube mince de 2 cm de section entre les lèvres. Pendant les époques de contrôle les sujets sur le dos gardaient simplement ce tube entre les lèvres, sans le serrer, et ils fixaient le centre de l'écran. Pendant les époques de tâche ils prononçaient le [u] en enserrant le tube entre les lèvres, également sans phonation.

Pendant le troisième scan fonctionnel nous avons testé l'attente de la perception de la distinction entre les deux [u]. Pendant les époques de tâche, l'instruction était pour le sujet de se préparer à entendre une séquence aléatoire des deux [u], normal et compensé, qu'il venait de prononcer, pour les distinguer (il était convenu qu'il ne recevrait pas effectivement après cette période d'attente cette séquence sonore). Pendant les époques de contrôle, les sujets étaient au repos et fixaient un point au centre de l'écran.

2.2. *Acquisition des données*

Les examens IRMf ont été effectués dans un imageur clinique Philips NT, 1.5 Tesla, équipé avec des techniques d'imagerie rapide de type EPI. Nous avons mesuré un volume cérébral composé de 26 coupes de 5 mm chacune. Ce volume a été centré sur la limite supérieure du thalamus et a été orienté parallèlement au plan bi-commissural CA-CP. Il a été mesuré 14 fois pendant chaque époque de contrôle et de tâche. La durée totale du scan était de 9'34''.

Une séquence en écho de gradient a été appliquée. Les principaux paramètres d'acquisition ont été : TR = 4560 ms, TE = 45 ms, angle de basculement = 90°, champ de vision = 256x256 mm2, matrice d'acquisition = 64x64, matrice de reconstruction = 128x128.

Le positionnement du volume cérébral exploré pendant les scans fonctionnels a été effectué sur base d'une séquence de repérage dans le plan sagittal, acquise en début d'examen. En fin d'examen, nous avons effectué un scan anatomique générant des images, à haute résolution spatiale, du volume cérébral mesuré.

2.3. *Traitement des données*

Pour le traitement des données nous avons utilisé le logiciel SPM (Statistical Parametric Mapping [Fri94]). Le traitement de chaque scan fonctionnel a comporté les étapes suivantes. La série des volumes fonctionnels a été d'abord réalignée par recalage sur le premier volume acquis. Après ce recalage, les volumes ont été déformés ("normalisés") à l'aide d'une transformation non-linéaire afin qu'ils puissent se confondre avec un volume de référence ("template") représenté dans un espace standard, typiquement l'espace MNI (du Montreal Neurologic Institute). Une transformation linéaire appliquée ensuite nous a permis d'obtenir les coordonnées dans l'espace de Talairach à partir des coordonnées dans l'espace MNI. Après lissage spatial de ces volumes "normalisés", le modèle linéaire général (GLM, general linear model) a été appliqué en vue de pouvoir effectuer les tests univariés (ANCOVA) au niveau de chaque voxel. Les tests statistiques nous ont permis d'obtenir les cartes fonctionnelles "SPM".

Le modèle linéaire général est une équation qui exprime les évolutions temporelles des pixels en une combinaison linéaire des facteurs explicatifs (facteurs d'intérêt et de non-intérêt), auxquels se rajoute un terme d'erreur supposé fluctuer normalement. Les contributions relatives de chacun de ces facteurs ont été déterminées à l'aide du calcul des moindres carrés. Des inférences quant à ces contributions ont été basées sur des statistiques de type F (F = 0.05). Les inférences finales pour le volume analysé ont été basées sur la probabilité de détecter un nombre minimum de clusters comprenant chacun au moins k voxels (k>5 voxels) présentant un score z au-delà d'un seuil préétabli (z>2.5).

3. RESULTATS

Nous avons visualisé les réponses fonctionnelles obtenues chez les sept sujets, sur la base d'une analyse de groupe. Les activations significatives obtenues (les "clusters") ont d'abord été représentées en couleur sur un "template" 3D. Pour des raisons de reproduction, ces régions activées, sont données ici en projection sur des coupes transversales (2D).

3.1. *Articulation normale du [u]*

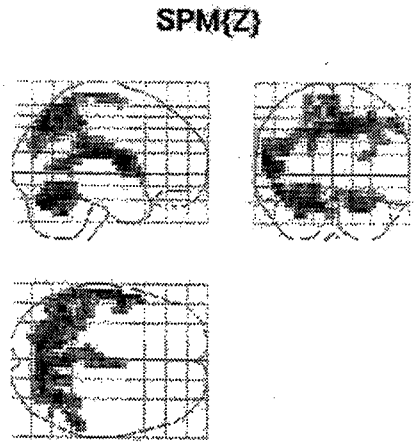


Figure 1

La Figure 1 ci-dessus représente les résultats en mode projectif, sagittal (en haut à gauche ; vue de droite), puis coronal (à droite ; vue d'arrière) et transverse (en dessous ; l'arrière est à gauche).

Pour chaque cluster, nous avons identifié les régions activées qui correspondent aux voxels les plus significatifs. Elles sont les suivantes.

Cluster 1 : Une représentation 3D de la surface du cortex montre clairement une activation périsylvienne gauche (ici bien visible de droite en projection sagittale). Dans ce cluster, le voxel le plus significatif est situé dans le *cortex prémoteur (6 BA) gauche*, tandis que les deux suivants sont localisés dans le *gyrus supramarginal (40 BA) gauche*.

Cluster 2 : les deux voxels les plus significatifs correspondent au *précuneus (7 BA) droit* (nettement visible en projection coronale), tandis que le troisième correspond au *gyrus supramarginal (40 BA) droit*.

Cluster 3 : les deux voxels les plus significatifs correspondent au *cervelet gauche* (visible en projection coronale).

Cluster 4 : les quatre voxels représentés correspondent au *cervelet droit* (cf. la projection coronale).

3.2. Articulation compensée du [u]

La Figure 2 ci-dessous représente les résultats en mode projectif. Pour chaque cluster, les régions activées correspondant aux voxels les plus significatifs sont les suivantes.

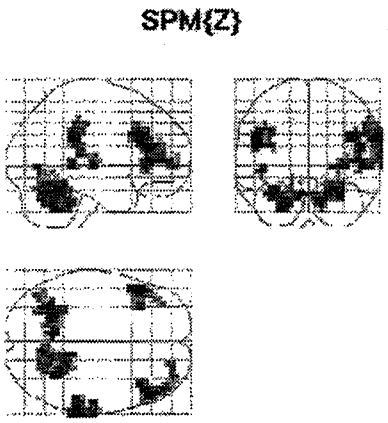


Figure 2

Cluster 1 : les trois voxels représentés correspondent au *cortex préfrontal frontal (8, 9, 46 BA) gauche.*

Cluster 2 : les deux premiers voxels correspondent au *cortex préfrontal (9, 46 BA) droit*, tandis que le dernier correspond au *gyrus frontal inférieur (45 BA) droit.*

Cluster 3 : le premier et le troisième voxel correspondent au *cervelet droit*, tandis que le deuxième correspond au *cervelet gauche.*

Cluster 4 : le premier voxel correspond au *gyrus supramarginal (40 BA) droit*, le deuxième *au gyrus temporal postéro-supérieur droit* (22 *BA*) et le troisième au *gyrus temporal moyen (21 BA) droit.*

3.3. Attente perceptive des [u] normal/compensés

La Figure 3 ci-dessous représente, en mode projectif, les résultats.

Pour chaque cluster, les régions activées correspondant aux voxels les plus significatifs sont les suivantes.

Cluster 1 : le voxel le plus significatif correspond au *cortex prémoteur mésial* (aire motrice supplémentaire, SMA, *6 BA* mésial) gauche.

Cluster 2 : les voxels correspondent au *cervelet droit.*

Cluster 3 : les deux premiers voxels représentés correspondent au *cortex préfrontal gauche (46 BA)* et le troisième au *gyrus frontal inférieur gauche (47 BA).*

Cluster 4 : les deux premiers voxels représentés correspondent au *cortex préfrontal droit (46 BA)* et le troisième au *gyrus frontal inférieur droit (47 BA).*

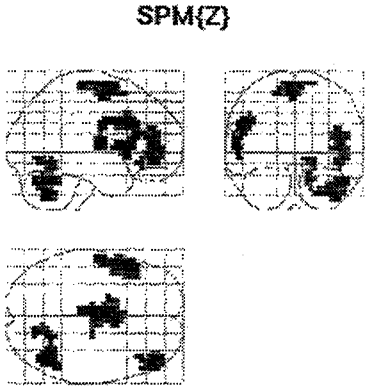


Figure 3

4. DISCUSSION ET CONCLUSIONS

La différence d'ensemble entre les conditions d'articulé normale et compensée fait apparaître clairement un changement de dominance hémisphérique: *l'équivalence correspond donc à un changement de dominance.* Contrairement à la dominance périsylvienne classiquement gauche de nos sujets tous droitiers, la tâche compensée recrute majoritairement à droite.

Rappelons d'emblée que cette tâche de prononciation de la voyelle [u] est bien dès le départ une tâche phonologique (pas un Ouh!). En témoigne à gauche l'activation de *40 BA* réputée fonctionner pour cet encodage phonologique [Pau93]. En fait il faudrait voir dans cette aire *40 BA* l'encodage du *but* articulatoire *global.* Soit la représentation — en relation avec les modalités perceptives corticalement voisines (proprioceptives, auditives, voire visuelles) — de l'effet phonémique à réaliser, sans que les moyens pour y parvenir par une action précise des effecteurs soient pris en compte dans cette zone. C'est du moins ce que l'on peut proposer depuis les travaux d'Hyvarinen sur le pariétal, jusqu'à ceux menés actuellement dans l'équipe de Jeannerod sur la préhension, en passant par Abbs pour la parole [Abb86]. Selon cette conception, cette aire *40 BA* est en relation avec le prémoteur (*6 BA*), en ce qu'il représenterait l'encodage moteur de ce but [Wis97]. (Notons que l'articulation statique est une action sans doute trop peu énergétique pour activer dans l'aire motrice primaire les parties orofaciales mises en jeu.) Toujours pour ce [u] non compensé, *40 BA* est aussi dans une moindre mesure activée à droite. Les zones du cervelet activées à gauche et aussi à droite correspondraient aux processus nécessaires pour la mise en œuvre d'un modèle interne, selon les propositions déjà anciennes d'Ito, reprises tout récemment [Ito00]. On peut se demander pourquoi l'activité du précuneus droit *(7 BA)*, n'apparaît que pour cette tâche. Une interprétation serait que c'est seulement pour cette activité motrice du [u] prototypique, stocké de longue date, qu'on fait appel à une "mémoire du corps" (depuis [Ber95]), ici mémoire proprioceptive de la parole.

Pour la tâche du [u] compensé, la dominance est à droite. En contraste avec le [u] non compensé, *40 BA* est ici

activée uniquement à droite. Cette aire est en relation avec une homologue droite de Broca *45 BA* (*pars triangularis*), laquelle est réputée, à gauche, à la fois pour la sémantique *et* la syntaxe des actions, y compris linguistiques ; alors que les autres aires de Broca *47 BA* (*pars orbitalis*) et *44 BA* (*pars opercularis*) sont plutôt spécialisées chacune respectivement, soit dans la sémantique, soit dans la syntaxe ([Dap99], [Pol99]). A droite toujours, pour *22 BA* ou l'homologue de Wernicke, qui pourrait être recrutée pour un *monitoring* ou contrôle du son qui correspondrait à l'action articulatoire compensée. Le rôle de *21 BA* droite est moins clair qu'à gauche, bien qu'elle puisse (selon [Hag99], p. 12) être aussi impliquée "in some form of phonological processing, either of the spoken input or the to-be-spoken output". L'activité du cervelet droit comme gauche correspondrait là aussi à la présence d'un modèle interne. Enfin le cortex préfrontal ou dorso-latéral préfrontal, droit et gauche, serait impliqué dans les tâches qui font appel à une mémoire de travail concernant l'attention pour l'action, ici l'équivalence motrice. Et ce sera le cas dans une tâche de perception, elle aussi difficile, comme celle qui va suivre.

La dernière tâche est inspirée d'une hypothèse de Catherine Best *et al.* (expérience non publiée) qui rend compte des activités cérébrales gauches et droites dans la catégorisation et la sous-catégorisation des sons linguistiques. Mais notons tout d'abord que l'activité observée pour l'aire motrice supplémentaire (ou SMA) peut être due, dans cette dernière tâche, à l'attente d'un *lancement* (ou *initiation*) d'une *séquence* de [u] compensés et non compensés à distinguer. Rappelons que l'idée principale de Best *et al* capitalise le fait que les sons non-linguistiques sont plutôt traités à droite et les catégories phonémiques à gauche. Partant, les distinctions à l'intérieur des catégories ou distinctions sous-catégorielles — effectuées à l'écoute d'une langue étrangère, ou à l'intérieur de sa propre langue dans le cas de variantes dialectales, par exemple — ces distinctions recruteraient aussi les aires droites homologues des aires du langage dans l'hémisphère gauche. Et c'est bien ce que Best *et al.* ont observé pour Broca et son homologue à droite dans une tâche de *perception*. Dans notre cas on peut considérer que l'activation gauche et droite de *47 BA* correspondrait à cette activité d'évaluation, en quelque sorte perceptivo-motrice, de la qualité du [u] compensé comme une sous-catégorie de [u]. Du moins c'est là notre hypothèse: un renforcement de l'activité des homologues du langage dans l'hémisphère droit qui pourrait rendre compte de la dominance observée, pour la *perception* et encore plus, dans la tâche précédente, pour la *production* d'une nuance dans une catégorie phonémique, ici une équivalence motrice.

Pour finir nous mentionnerons que ce résultat majeur d'une dominance hémisphérique droite chez des sujets qui sont classiquement dominants à gauche pour le langage, dans le cas d'un recodage d'une variante équifinale en son, sera d'autant moins surprenant que nous avons déjà obtenu cette dominance droite pour la même équivalence motrice dans une première expérience (non publiée) en IRMf Fast Field Echo, chez 7 sujets, dont 5 identiques à ceux de cette imagerie écho planaire (EPI).

Remerciements : A nos 7 sujets entraînés (Pierre Badin, Louis-Jean Boë, Albert Rilliard, Solange Rossato, Christophe Savariaux, Jean-Luc Schwartz, Anne Vilain ; et aux volontaires précédents : Yohan Payan et Frédérique Sannier). A Pierre Badin pour son aide dans l'analyse des IRM anatomiques des performances normales et compensatrices des sujets et le calcul des fonctions d'aires pour obtenir l'équivalence des fonctions de transfert. A Catherine Best pour l'illumination qu'elle nous a procurée en perception, que nous avons pu reprendre en production et confirmer *in fine* en perception.

BIBLIOGRAPHIE

[Abb86] Abbs J. (1986) Invariance and variability in speech production: A distinction between linguistic intent and its neuromotor implementation, in Perkell J., Klatt D., Invariance and variability in speech processes, pp. 203-219.

[Ber95] Berthoz A., *et al.* (1995) Spatial memory of body linear displacemeny: What is beeing stored?, Science 269, pp. 95-98.

[Dap99] Dapretto M., Bookheimer S.Y. (1999) Form and content: Dissociating syntax and semantics in sentence comprehension, Neuron 24(2), pp.427-432.

[Fri94] Friston K.J. (1994) Statistical Parametric Mapping.Functional neuroimaging, pp. 79-93.

[Hag99] Hagoort P., *et al* (1999) The neural circuitry involved in the reading of German words and pseudowords: A PET Study, J. of Cognitive Neuroscience 11, pp. 383-398.

[Ito00] Ito M. (2000) Internal model visualized,. Nature 403, pp. 153-154.

[Pau93] Paulesu E., *et al* (1993) The neural correlates of the verbal component of working memory",. Nature, 362, pp. 342-345.

[Pol99]] Poldrack R.A., *et al* (1999) Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex, NeuroImage 10, pp. 15-35.

[Sav95] Savariaux C., *et al.* (1995) Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube : A study of the control space in speech production, JASA 98(5), pp. 2428-2442.

[Sav99] Savariaux C., *et al.* (1999) Compensation strategies for the perturbation of French [u] using a lip tube: II. Perceptual analysis, JASA 106(1), pp. 381-393.

[Wis97] Wise, S.P., *et al.* (1997) Premotor and parietal cortex: Corticocortical connectivity and combinatorial computations, Annual Review of Neurosciences 20, pp. 25-42.

Prosodie/phonologie

Le registre en voix parlée : un indicateur social pour homme seulement ?

Monique Demers

Département des arts et lettres – Université du Québec à Chicoutimi, CANADA
Tél. : 418-545-5011, poste 5254 – Télécopieur : 418-545-5012
Courriel : mdemers@uqac.uquebec.ca

ABSTRACT

The notion of register here corresponds to that of the voice F_0 level (mean F_0) as well as that of the F_0 range (F_0 max - F_0 min). These parameters are studied through a spontaneous corpus of male and female voices distributed among three socioprofessionnal status (low, mid, high) and among two varieties of French (from France and from Québec). First, it is observed that register is an indicator of social status among males : a low register voice and a wide F_0 range have a tendency to characterize speakers with a high status. Second, the register also seems to be a geographic indicator among males : the voice of French speakers is generally not as low, but the F_0 range is wider than the voice of Québec speakers. The vocal register of females with a different socioprofessional status and a different geographic origin does not show any significant difference.

1. INTRODUCTION

Le registre vocal des hommes et des femmes a été étudié selon des points de vue physiologiques, artistiques ou sociolinguistiques, notamment par Luchsinger et Arnold [Luc65], Fitch et Holbrook [Fit70], Linke [Lin73], Fournier [Fou94][, Boë *et al.* [Boë75], Graddol et Swann [Gra89], Henton [Hen89]. Outre les différences physiologiques évidentes de hauteur entre une voix d'homme et une voix de femme, il apparaît régulièrement que cette différence est moins élevée que ce que permettrait d'attendre la voix physiologique, ou naturelle (une octave). Et selon de Pinto et Hollien [Pin82], plus on avance dans le siècle, plus la voix des femmes aurait tendance à s'abaisser. De plus, malgré la capacité physiologique d'étendue des voix de femme (cordes vocales plus courtes, plus minces, plus flexibles), en principe supérieure à celle des voix d'homme, il est observé que l'étendue, ou la variation, des voix de femme est souvent comparable, voire inférieure, à l'étendue des voix d'homme. Ainsi les différences entre les voix d'homme et de femme paraissent-elles relever non seulement de la voix physiologique, mais aussi d'une voix *sociale*.

Cela dit, on peut se demander jusqu'où porte cette voix *sociale*, aussi bien chez les hommes que chez les femmes. Dans quelle mesure des facteurs sociaux tels le statut socioprofessionnel et l'origine géographique ont-il un effet sur le registre vocal des hommes et des femmes ? Et cet effet est-il le même sur les voix d'homme que sur les voix de femme ?

1.1 Le registre et le statut socioprofessionnel

Les caractéristiques vocales apparaissent comme des repères significativement fiables pour l'identification du statut social, du moins chez les hommes. Par exemple, l'étude de Brown et Lambert [Bro76] montre que le statut social d'hommes parlant le français du Canada est identifié aussi bien par des pairs que par des anglophones ne parlant pas français.

Les explications éthologiques des corrélations entre la hauteur de la voix et les traits de personnalité données par Ohala [Oha83] ont été marquantes. Selon Ohala, la voix du socialement *fort*, celui qui donne l'apparence du *grand*, de la *domination* est grave ; à l'inverse, la voix du socialement *faible*, du *petit*, du *subordonné* est plus aiguë. Il s'agirait là d'un code fréquenciel inné (« frequency code »). Il n'est pas fait mention de l'étendue (différence entre F_0 max et F_0 min). Toutefois, des études perceptuelles ont régulièrement associé une étendue, ou une variation importante, à l'attribution de traits de personnalité positifs (notamment [Bro74, Bez88]).

1.2 Le registre et l'origine géographique

La question des différences prosodiques, particulièrement la différence d'étendue entre les voix d'homme français et québécois, a été étudiée par les phonéticiens québécois il y a déjà quelques décennies. Toutefois, ces premières études [Gen66 ; Bou68 ; Hol68] présentent une base descriptive limitée (de 3 à 7 locuteurs, presque toujours des hommes, cultivés, la plupart du temps en contexte de lecture). De plus, une certaine idéologie amenait alors à considérer le français du Québec comme un sous-produit du français de France. Quoi qu'il en soit, la caractéristique prosodique distinctive la plus évoquée est toujours celle de la mélodie de la phrase : le français québécois serait « monotone » alors que le français hexagonal serait plus « chantant ». Il faut pourtant souligner que cette caractéristique relève davantage d'une impression que du résultat d'analyses quantitatives. Les résultats obtenus par Boudreault [Bou68] et Holder [Hol68] les amènent d'ailleurs à poser des hypothèses autres qu'une étendue réduite pour expliquer cette monotonie (caprices individuels, contexte ou haut niveau culturel, timbres vocaliques plus clairs, etc.). Par ailleurs, une étude récente [Bis00], faite à partir de la lecture de bulletins de nouvelles, tend à montrer que la variété québécoise aurait une étendue plus importante que la variété française. En ce qui concerne les voix de femme, aucune étude ne permet véritablement de comparer la voix des Françaises avec celle des Québécoises.

2. MÉTHODOLOGIE

Un sommaire de la méthodologie est ici présenté. On trouvera une description détaillée dans [Dem00].

2.1 Corpus

L'étude porte sur 30 voix d'homme et 30 voix de femme en période de vie active (15 Québécois, 15 Français ; 15 Québécoises, 15 Françaises). Chaque groupe de locuteurs est représentatif de trois milieux sociaux (faible, moyen, élevé). Le corpus québécois est issu du *Corpus Montréal 84* [Thi90] et le corpus français, du *Corpus Paris 97* (recueilli par l'auteure). L'analyse est issue d'extraits d'entrevues semi-dirigées portant sur l'emploi, extraits choisis à partir de la qualité acoustique et du contenu informatif. Pour chacun des locuteurs, les séquences analysées totalisent au moins 60 s, comportent une moyenne de 3388 valeurs de F_0, de 10 énoncés d'une durée moyenne de 5,91 s.

2.2 Traitement instrumental

L'analyse acoustique a été réalisée à partir du logiciel de traitement de parole *Computerized Speech Lab* (CSL) de Kay Elemetrics Corp. Le réglage des paramètres d'analyse a été fixé selon les caractéristiques de chaque locuteur. Il arrive que l'extraction automatique de F_0 par CSL produise des valeurs insolites. Toutes les valeurs ont donc été révisées puis ajustées au besoin à partir des valeurs environnantes ou de la durée de la période.

2.3 Traitement statistique

Les trois catégories principales de mesures statistiques suggérées par Jassem [Jas71] pour l'analyse de la hauteur et de l'étendue de la voix sont : (i) comme mesure de tendance centrale, la moyenne arithmétique des valeurs de F_0 ; (ii) comme mesures de distribution, le degré d'asymétrie de la courbe des fréquences (*skewness*), qui renseigne sur les fréquences les plus utilisées, ainsi que le degré d'aplatissement (*kurtosis*), qui informe sur la diversité des fréquences utilisées (non significative, cette dernière mesure n'est pas présentée) ; (iii) comme mesure d'étendue, l'étendue estimée à partir de l'écart type. La présente analyse utilise chacune de ces catégories de mesure avec quelques adaptations et ajouts.

Compte tenu que la perception de F_0 s'effectue selon une loi logarithmique, les valeurs hertziennes sont transformées en valeurs tonales selon la formule conventionnelle suivante :

(1) F_0 moyenne (en tons) = $19{,}93 * \log_{10}(F_0 \text{ moy} / 16{,}35 \text{ Hz})$

L'étendue estimée (96% des fréquences laryngiennes) est calculée à partir de l'écart type selon la formule suivante :

(2) Étendue estimée = F_0 moy ± (2*écart type)

La conversion en tons se fait ensuite à partir de la formule (3) :

(3) Étendue tonale estimée = $19{,}93 * \log_{10}(F_0 \text{ max} / F_0 \text{ min})$

L'étendue mesurée est aussi calculée afin de pouvoir comparer avec des études antérieures. En effet, si les études sur l'anglais et sur le français européen considèrent presque toujours l'étendue estimée, celles sur le français québécois sont réalisées uniquement à partir de l'étendue mesurée.

(4) Étendue mesurée = F_0 max produite - F_0 min produite

La conversion en tons se fait à partir de la formule (3).

Des analyses de variance à trois facteurs (statut social, sexe, origine géographique) sont effectuées pour déterminer l'effet de ces facteurs sur les variables prosodiques retenues. Les hypothèses de base sous-jacentes sont rencontrées, c'est-à-dire la normalité et l'homogénéité des variances.

3. RESULTATS ET DISCUSSION

Les résultats de la présente étude sur le registre portent donc sur la hauteur de F_0 et sur l'étendue. Ils sont toujours présentés en relation avec le sexe auquel s'ajoutent le statut socioprofessionnel puis l'origine géographique. Ces résultats sont discutés au fur et à mesure de leur présentation.

3.1 Le registre et le sexe

Comme dans la plupart des études des dernières décennies, les caractéristiques distinctives des voix d'homme et des voix de femme ici observées ne correspondent pas à ce que les seules différences physiologiques permettraient d'attendre (*Table 1*).

Table 1. Résultats pour le registre comme indicateur du sexe. Légende : H = homme ; F = femme. *** = $p < 0{,}001$

SEXE	HAUTEUR		ETENDUE	
	Moyenne tonale	Coefficient d'asymétrie	Mesurée (en tons)	Estimée (en tons)
H/F	16,45/21,13 ***	0,93/0,55 ns	7,81/7,38 ns	6,65/5,91 ns
Différences interindividuelles				
H/F	4,96/3,16	3,53/2,93	7,88/7,13	8,35/5,56

En effet, la différence de moyenne entre la hauteur tonale des 30 voix d'homme et des 30 voix de femme est de 4,68 tons, ce qui est inférieur à l'octave (6 tons) attendue. De même, alors qu'une étendue plus grande est prévisible chez les femmes, les données ne permettent d'observer aucune différence significative. L'examen des différences interindividuelles (i) par les valeurs extrêmes montre un écart toujours plus important entre les hommes qu'entre les femmes ; (ii) l'examen par classe moyenne (± 1ton) va dans le même sens : plus d'hommes (13 pour la hauteur et 16 pour l'étendue estimée) que de femmes (respectivement, 9 et 13) s'écartent de cette classe moyenne.

Discussion. D'une part, la différence de moyenne entre les deux sexes montre que la voix *sociale* des femmes tend à se rapprocher du modèle vocal masculin : diminution de la hauteur vocale moyenne et diminution de l'étendue de la voix. L'interprétation la plus courante de cette voix *sociale*, ici désignée comme *masculine* et *féminine*, tourne autour du fait qu'une société androcentrique (qui a tendance à privilégier le groupe des hommes) aurait rendu péjoratives les caractéristiques naturelles des voix de femme : une voix aiguë, *swoopy*, émotive, etc. [Hen89]. Pour des raisons sans doute apparentées, la radio et la télévision proposent un modèle de voix féminine calqué sur le modèle masculin : une voix plutôt

basse, sans trop de variation (notamment [Lin73 ; Hen95]). Les analyses perceptuelles qui ont mesuré l'effet de ces tendances vocales féminines sont, à ma connaissance, peu nombreuses et datent déjà de quelques décennies ([Luc68] ;[Lin73]) ; ces dernières montrent toutefois qu'il n'y a pas nécessairement de corrélation entre une F_0 plus basse et l'« efficacité » de la voix. Pronovost [Pro42] avait déjà suggéré que le registre usuel (le plus souvent utilisé) est efficace lorsqu'il coïncide à peu près avec le registre naturel (mesuré à partir de la voix chantée). Plus récemment, et dans un contexte bien spécifique, Smith [Smi92] observe que les femmes japonaises en situation d'autorité tentent plutôt de s'approcher du modèle enfantin. Des études perceptuelles sur le sujet sont incontestablement requises.
D'autre part, les différences interindividuelles, toujours plus importantes entre les voix d'homme qu'entre les voix de femme soulèvent la question du pourquoi. Deux autres facteurs sociaux sont ici examinés.

3.2 Le registre, le sexe et le statut

Trois statuts socioprofessionnels ont été analysés (faible, moyen et élevé). Cependant, les écarts les plus significatifs se retrouvant entre les groupes extrêmes (faible et élevé), seuls ceux-ci seront présentés. Les résultats obtenus confirment que le registre vocal des locuteurs de statut socioprofessionnel faible se distingue de celui des locuteurs de statut socioprofessionnel élevé, du moins chez les hommes (*Table 2*).

Table 2. Résultats pour le registre comme indicateur du statut socioprofessionnel (faible/élevé) chez les hommes et chez les femmes. Légende : f = faible ; é = élevé ; H = homme ; F = femme. * = p < 0,05 ; ** = p < 0,01

SEXE * STATUT	HAUTEUR		ETENDUE	
	Moyenne tonale	Coefficient d'asymétrie	Mesurée (en tons)	Estimée (en tons)
H f/é	17,14/15,71 *	1,45/0,26 **	ns	6,10/7,48 *
F f/é	ns	ns	ns	ns

Dans le corpus analysé, il apparaît que les voix d'homme de statut socioprofessionnel faible ont une moyenne tonale plus élevée (voix plus haute), une distribution de fréquences plus asymétrique (i.e. une plus grande utilisation de fréquences basses par rapport à la valeur moyenne absolue de leur sous-groupe, 120,90 Hz) et une étendue plus étroite que les locuteurs de statut socioprofessionnel élevé. Chez les femmes, aucun des paramètres prosodiques étudiés ne permet de distinguer le statut socioprofessionnel.
Discussion. Dans un premier temps, les différences de hauteur observées chez les hommes de statut faible et élevé vont dans le sens des observations faites à ce jour, à partir de voix d'homme. Faudrait-il voir dans la hauteur vocale le paramètre sociobiologique par excellence, c'est-à-dire celui par lequel on distingue non seulement les hommes des femmes, mais aussi les *forts* des *faibles*. Le moyen prosodique privilégié pour affirmer masculinité (voir [Ter66] sur les voix efféminées) et pouvoir ? Quant à l'étendue, quelques études perceptuelles réalisées à partir de voix d'hommes établissent aussi un lien entre la voix qui utilise une étendue large et le milieu socioprofessionnel élevé, notamment [Bez88]. Mais comment expliquer que les femmes entre elles semblent échapper à ce stéréotype vocal de la *plus forte* ? Il est généralement admis que le langage des femmes tend à se rapprocher de la forme standard [Lab66 ; Tru74 ; Tho75]. Il en découlerait que la femme de statut faible –en tout cas plus que l'homme de statut faible– est elle aussi soucieuse de la bonne prononciation, du bon mot, de la bonne construction de phrase et du *bon ton*, ce qui nivellerait les différences interfemmes. On peut toutefois se demander ce qu'il adviendrait en contexte plus formel. La femme de statut élevé ne pourrait-elle jouer de la *carte vocale* qu'en situation formelle ? Lucci [Luc83] a observé, au contraire, que plus le contexte est informel, plus la voix devient grave. Les observations proviennent cependant d'un corpus limité (3 femmes de statut élevé, en contexte de lecture, de conférence et d'interview), qui n'a pas permis d'analyses statistiques. La question mériterait qu'on y revienne.

3.3 Le registre, le sexe et l'origine géographique

On se rappelle que la plupart des données disponibles sur la comparaison des voix françaises et des voix québécoises portent sur l'étendue vocale des hommes. Dans la *Table 3* sont comparées la hauteur et l'étendue des voix d'homme et de femme françaises et québécoises.

Table 3. Résultats pour le registre comme indicateur de l'origine géographique (France/Québec) chez les hommes et chez les femmes. Légende : H = homme ; F = femme ; Fr = France ; Q = Québec. * = p < 0,05 ; ** = p < 0,01

SEXE* ORIGINE	HAUTEUR		ETENDUE	
	Moyenne tonale	Coefficient d'asymétrie	Mesurée (en tons)	Estimée (en tons)
H Fr/Q	17,10/15,85 **	1,32/0,51 *	ns	7,22/5,75 **
F Fr/Q	ns	ns	ns	ns

Il apparaît que le registre peut aussi être un indicateur de l'origine géographique chez les hommes. En effet, les hommes français ont une voix plus haute (conforme à l'impression toujours véhiculée), font une utilisation plus asymétrique de l'ensemble des fréquences et ont une étendue plus large que les hommes québécois (ce qui est contraire aux résultats des études antérieures en contexte de lecture, mais peut correspondre à l'impression de « monotonie »). Chez les femmes (peu ou pas d'études antérieures disponibles), on n'observe aucune différence significative entre le registre vocal des Françaises et celui des Québécoises.
Discussion. Pour des raisons historiques évidentes (éloignement géographique puis au moment de la Conquête par les Anglais, coupure totale avec la France), le français du Québec s'est développé différemment de celui de la France (tant du point de vue phonétique que lexical ou morphosyntaxique), aussi bien chez les hommes que chez les

femmes. Néanmoins, du point de vue du registre, seuls les hommes présentent des distinctions liées à l'origine géographique (les Québécois ont une voix plus basse –il semblerait qu'il s'agisse là d'une influence nord-américaine– et une étendue réduite). Pourquoi n'y a-t-il aucune de ces distinctions significatives chez les femmes ? Évidemment, on peut encore évoquer la tendance féminine à utiliser une langue plutôt *standard*, tendance qui inciterait les femmes à rester plus proche du modèle européen. Pourtant, des différences aux autres niveaux linguistiques sont notables. Les voies de spéculation sont multiples, mais surtout, les zones à explorer encore bien vastes.

4. CONCLUSION

Dans le corpus de parole spontanée ici analysé, la pratique prosodique masculine paraît utiliser le registre non seulement comme un indice distinctif du sexe, mais aussi comme un indice distinctif du statut social et de l'origine géographique tandis que la pratique prosodique féminine paraît plutôt utiliser le registre à des fins d'égalité, ou d'égalitarisme, aussi bien par rapport aux hommes que par rapport aux femmes qui ont des caractéristiques sociales différentes. Malheureusement, les résultats sur les voix de femme ne peuvent être comparés avec ceux d'études antérieures en raison de l'absence quasi totale d'études socioprosodiques du registre à partir de voix de femme, tant du point de vue de l'appartenance socioprofessionnelle que de l'origine géographique –France/Québec. Il faudra non seulement faire des analyses de la production en contextes situationnels variés, des analyses de la perception, mais aussi des analyses sur les fondements de cette perception.

BIBLIOGRAPHIE

[Bez88] van Bezooijen, R. (1988), "The relative importance of pronounciation, prosody, and voice quality for the attribution of social status and personality characteristics" R. van Hout et U. Knops (dir.), Languages attitudes in the Dutch language area, Dordrecht, Foris Publications, p. 85-103.

[Bis00] Bissonnette, S., Le registre en voix parlée, M. Demers (dir.), Québec, les Éditions Nota Bene, (à paraître).

[Boë75] Boë, L.-J., M. Contini et H. Rakotofiringa (1975), « Étude statistique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques du français », Phonetica, vol. 32, p. 1-23.

[Bou68] Boudreault, M. (1968), Rythme et mélodie de la phrase parlée en France et au Québec, Paris, Klincksieck et Québec, Presses de l'Université Laval, p. 101-122.

[Bro74] Brown, B.L., W.J. Strong et A.C. Rencher (1974), "Fifty-four voices from two: the effects of simultaneous manipulations of rate,mean fundamental frequency and variance of fundamental frequency on ratings of personality from speech", JASA, vol. 55, n° 2, p. 313-318.

[Bro76] Brown, B.L. et W.E. Lambert (1976), "A cross-cultural study of social status markers in speech", Canadian Journal of Behavioral Science, vol. 8, p. 39-55.

[Dem00] Demers, M. (2000), « La voix du plus *fort* », M. Demers (dir.), Le registre en voix parlée, Québec, Les Éditions Nota Bene, (à paraître).

[Fit70] Fitch, J.L. et A. Holbrook (1970), " Modal fundamental frequency of young adults ", Archives of Otolaryngology, vol. 92, p. 379-382.

[Fou94] Fournier, C. (1994), La voix, un art et un métier, Seyssel, Éditions Comp'Act, 316 p.

[Gen66] Gendron, J.-D. (1966), Tendances phonétiques du français parlé au Canada, Paris, Klincksieck et Québec, Presses de l'Université Laval, p. 152-161.

[Gra89] Graddol, D. et J. Swann (1989), Gender voices, "The voice of authority", Oxford, Basil Blackwell, p.12-40.

[Hen89] Henton, C. (1989), "Fact and fiction in the description of female and male pitch", Language and Communication, vol. 9, n° 4, p. 299-311.

[Hen95] Henton, C. (1995), "Pitch dynamism in female and male speech", Language and Communication, vol. 15, n° 1, p. 43-61.

[Hol68] Holder, M. (1968), « Étude sur l'intonation comparée de la phrase énonciative en français canadien et en français standard », P.R. Léon (dir.), Recherches sur la structure phonique du français canadien, Studia Phonetica, vol. 1, Montréal, Didier, p. 175-191.

[Jas71] Jassem, W. (1971), "Pitch and compass of the speaking voice", Journal of the International Phonetics Association, vol. 1, p. 59-68.

[Lab66] Labov, W. ([1966] 1982) The social stratification of English in New York City, Washington, Center for Applied Linguistics, 655 p.

[Lin73] Linke, C.E . (1973), "A study of pitch caharacteristics of female voices and their relationship to vocal effectiveness", Folia Phoniatrica, vol. p. 173-185.

[Luc65] Luchsinger, R. et G.E. Arnold (1965), Voice-Speech-Language, Belmont (Californie), Wadsworth Publishing Company.

[Luc83] Lucci, V. (1983), Étude phonétique du français contemporain à travers la variation situationnelle, Grenoble, Université des langues et des lettres de Grenoble, p. 135-166.

[Oha83] Ohala, J. (1983), "Cross-language use of pitch : an ethological view", Phonetica, vol. 41, p. 1-16.

[Pin82] de Pinto, O. et H. Hollien (1982), "Speaking fundamental frequency characteristics of Australian women : then and now", Journal of Phonetics, vol. 10, p. 367-375.

[Pro42] Pronovost, W. (1942), "An experimental study of methods for determining natural and habitual pitch," Speech Monographs, vol. 9, p.111-123.

[Smi92] Smith, J. (1992), "Women in charge : Politeness and directives in the speech of Japanese women", Language and Society, vol. 21, p. 59-82.

[Ter66] Terango, L. (1966), "Pitch and duration characteristics of the oral reading of males on a masculinity-feminity dimension", Journal of Speech and Hearing Research, vol. 9, p. 590-595.

[Thi90] Thibault, P. et D. Vincent (1990), Un corpus de français parlé : Montréal 1984, Québec, CIRAL, 145 p.

[Tho75] Thorne, B. et N. Henley (dir.) (1975), Language and sex : difference and dominance., Rowley, Mass., Newbury House Publishers, 311 p.

[Tru74] Trudgill, P. (1974), The Social Differenciation of English in Norwich, Cambridge, Cambridge University Press.

Différenciation prosodique précoce chez de jeunes enfants bilingues coréen-français

*Han Youmi & Jean-Yves Dommergues**

Université Paris 7 (*également Université Paris 8)
Laboratoire de Phonétique, 10 rue Charles V, 75004 Paris, France
Mél: agnesyumi@yahoo.com
dommerg@ccr.jussieu.fr

ABSTRACT

This present study investigates the acquisition of French prosody by two native Korean children aged 3 to 3;3 as they learn French as a second language. More specifically, it focuses on prosodic correlates of semantic categories in two-word utterances spoken in French by these children. The results suggest that two differentiated prosodic systems emerge quite early and rapidly in young bilinguals. By age 3 and within a 9-month period of regular contact with the French language, the prosodic competence of these bilingual children in French displayed native-like similarities with that of a French monolingual child used as a control. In particular, this study addresses the current issue of whether one or two linguistic systems are at work in young bilinguals.

1. INTRODUCTION

Plusieurs auteurs ont déjà suggéré, à l'instar d'Ingram [Ing81] ou de Paradis [Par&Gen96], que les premiers mots constituant les énoncés de jeunes enfants se différencient mutuellement par des variations prosodiques parfois subtiles. Mais peu de recherches ont à ce jour évalué les capacités de différenciation prosodique de jeunes bilingues. La présente étude, qui s'inscrit dans cette ligne de recherche, concerne plus spécifiquement la prosodie des énoncés à deux mots produits par de jeunes enfants bilingues coréen-français dont le français est la langue seconde. Les auteurs (cf. aussi [Han&Dom99]) tentent de répondre aux trois questions suivantes:

1) Peut-on dégager des corrélats prosodiques des catégories sémantiques dans les énoncés à deux mots chez ces enfants bilingues lors de leur production en français?

2) Ces bilingues présentent-ils des contours mélodiques similaires à ceux d'un enfant français monolingue lors de la production d'énoncés français à deux termes?

3) Si tel est le cas, cette maîtrise des patrons prosodiques français par les jeunes apprenants coréens peut-elle s'établir rapidement, par exemple en quelques mois, quand ces enfants vivent dans une situation bilingue?

2. MÉTHODE

2.1 Sujets

Ont participé à cette étude deux enfants coréens bilingues et un enfant français monolingue qui joue le rôle de sujet contrôle. Les deux premiers sont nés en Corée et, à l'âge de 3 mois, s'installent à Paris avec leurs parents. Ces derniers décident de scolariser leurs enfants dès l'âge de 2 ans et demi environ (2;6 ans). C'est donc à cet âge qu'ils commencent à apprendre le français. Ils parlent français à l'école, et coréen à la maison. Au début de l'observation, ils ont respectivement 3 ans (C-GH) et 3;3 ans (C-DY). Quant à l'enfant français (F-JH), c'est un parisien monolingue de 2;2ans.

2.2 Procédure et choix du corpus

Le corpus est un ensemble d'énoncés à deux termes produits par deux enfants coréens bilingues et un enfant français. Nous ne nous intéressons ici qu'à la production en français. Chaque enregistrement a été effectué en milieu familial, au cours d'un entretien d'environ une heure. La conversation, pilotée par le premier auteur, s'est articulée autour de questions du genre "qu'est-ce que c'est?", "raconte-moi ce que tu vois dans ce livre" etc. Il était attendu de l'enfant des productions (souvent des réponses) spontanées. Nous avons éliminé du corpus toutes les réponses en imitation ou en écho, pour ne conserver que les seules réponses ou explications spontanées. Nous n'étudions donc ici que des énoncés produits par l'enfant de manière autonome et sur un mode affirmatif.

Ce corpus a été obtenu sur une période de deux mois pour chaque sujet. Si l'on compare le nombre de mots prononcés par chaque locuteur, on constate un certain déséquilibre entre ces enfants; mais pour notre étude, nous nous avons exploité les seuls énoncés à deux termes, le critère étant la longueur moyenne des énoncés (désormais MLU): MLU= 2,16 mots en moyenne pour les trois enfants); les valeurs

individuelles correspondantes sont les suivantes: 2,1 mots (Français-JH, âgé de 2;2 ans), 1,83 mot (Coréen-GH, âgé de 3 ans) et 2,69 mots (Coréen-DY, âgé de 3;3 ans). La figure 1 donne la distribution globale des quatre types d'énoncés constituant le corpus initial.

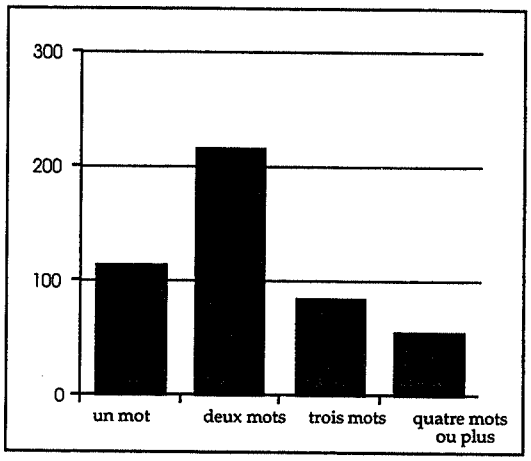


Figure1: Distribution des types d'énoncés: un mot, deux mots, trois mots et quatre mots ou plus. Corpus global des trois enfants (N = 496 énoncés).

La figure 1 montre que les énoncés à deux mots sont les plus fréquents chez ces enfants malgré les différences d'âge: c'est le corpus finalement retenu. Le nombre total d'énoncés à deux mots s'élève à 261 sur 496 énoncés effectivement recueillis (53 %): 110 par l'enfant français (43% de ses énoncés), 19 (52%) par C-GH et 132 (63%) par C-DY.

3. ANALYSE DU CORPUS

Le corpus retenu de 261 énoncés à deux mots est analysé du double point de vue de la sémantique et de la prosodie.

3.1 Catégorisation sémantique

Lorsque les enfants relient deux termes dans un énoncé, cela représente la première opération de prédication et manifeste la mise en œuvre de "notions cognitives" reflétant ce qu'ils sont en train d'apprendre du monde [Blo & Lah78]. De fait, nous avons classé ces énoncés en six catégories sémantiques:

- Catégorie 1 (désormais C1): agent-action (ex: "papa-parti", "bébé- tombé").
- Catégorie 2 : démonstratif-entité (ex: "ça-bateau", "ça-ballon").
- Catégorie 3 : agent-objet (ex: "papa -chapeau", "papa–pantalon").
- Catégorie 4: entité-qualificatif (ex: "ballon-rouge", "pantalon-bleu", "grand- bateau").
- Catégorie 5 : action-objet (ex: "parti-voiture", "joue-ballon").
- Catégorie 6 : action-localisation (ex: "tombé-là", "partir-là"). Pour ces trois enfants, la proportion globale de ces différentes catégories s'est avérée similaire.

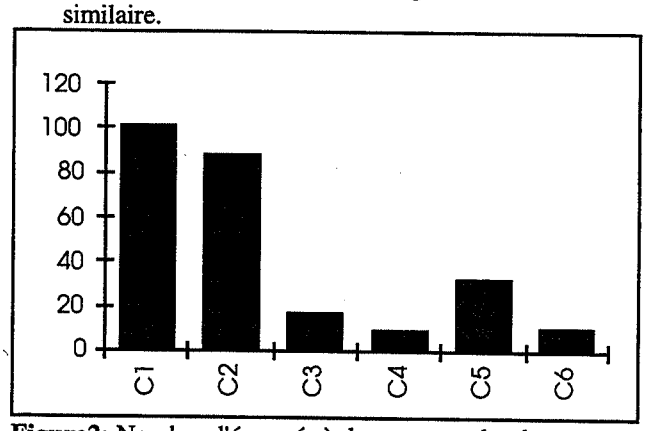


Figure2: Nombre d'énoncés à deux mots selon leur catégorisation sémantique (total N=261): six catégories.

La figure 2 montre que les catégories C 1 et C 2 sont les plus fréquemment utilisées par les enfants: ils représentent en effet 63 % du total des énoncés à deux mots.

3.2 Catégorisation prosodique

Les énoncés sont analysés grâce au logiciel "Signalyse" pour relever des invariants à partir des critères prosodiques F0 et Durée. Ces énoncés ont été divisés en deux parties, selon la présence ou l'absence de pause entre les deux mots (pause moyenne = 110 ms):

- **avec pause** (//) entre les deux mots, deux patrons apparaissent:

premier mot // deuxième mot
P1: F0 montant (+) // F0 descendant (-)
P2: F0 descendant (-) // F0 plat ou légèrement descendant (-)

- **sans pause** entre les deux mots:

premier mot, deuxième mot
P3: F0 montant (+), F0 descendant (-)

Ces patrons émergent de façon similaire chez les trois enfants. Notons que C1 a dû être sous-catégorisé en C1' et C1".

Dans le premier ensemble, avec pause interlexicale, quatre catégories (C1', C2, C5, C6) se réalisent avec le contour mélodique [F0 montant + // descendant -]: il s'agit du Patron P1. Mais C1 connaît une autre réalisation: [F0 descendant - // descendant -] (désormais C1"): il s'agit du Patron P2.

Dans le second ensemble, sans pause interlexicale, les catégories C3 et C4 se réalisent avec un seul contour

mélodique de type [F0 montant +, descendant-]: c'est le Patron P3. Nous obtenons donc finalement trois patrons essentiels de contours mélodiques à partir des énoncés à deux mots.

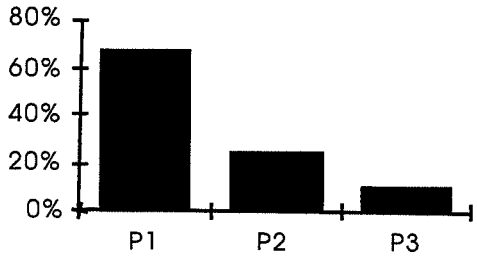


Figure3: Pourcentage d'énoncés à deux mots selon leur catégorisation prosodique (N = 261): patrons P1, P2 et P3

La figure 3 indique que P1 est globalement le plus utilisé par ces enfants (67%). Examinons maintenant comment les 6 catégories sémantiques et les trois patrons sont associés.

4. RÉSULTATS

4.1 Corrélats prosodiques des catégories sémantiques

L'examen parallèle des six catégories sémantiques et des trois patrons prosodiques a montré une forte association entre eux. La table 1 rend compte de cette association, c'est-à-dire des corrélats prosodiques des sous-catégories sémantiques.

Table 1: Corrélats prosodiques globaux des sous-catégories sémantiques.

patrons	P1	P2	P3
patrons prosodiques			
sous-catégories	C1' C2 C5 C6	C1"	C3 C4

La figure 4 représente la distribution des trois patrons prosodiques chez les trois enfants (F- est l'enfant français contrôle et les deux C- sont les enfants coréens).

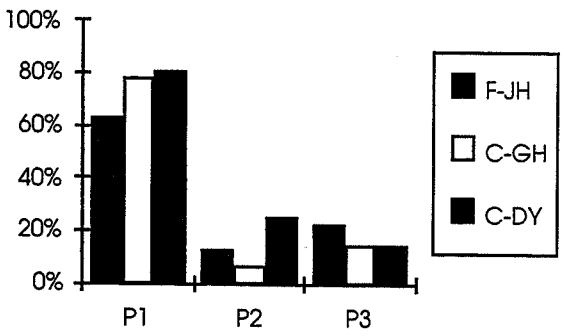


Figure4: Distribution des patrons prosodiques chez les trois enfants : deux bilingues C- et un monolingue F- (total N = 261).

Cette figure illustre la similarité des occurrences relatives des patrons P1, P2 et P3 chez les trois enfants de l'étude.

4.2 Similarités entre les deux bilingues et le monolingue français

4.2.1. Comparaison de la durée

Globalement (trois enfants pour les six catégories), la durée moyenne du deuxième mot (520ms) des énoncés à deux mots est plus longue que celle du premier (320 ms). Un *test-t* montre que cette différence est significative: t(40)= p<.0001.

Cette différence reste du même ordre chez les deux enfants bilingues mis ensemble (t(26)=p<.0001) et chez l'enfant monolingue (t(12)=p<.0001).

En revanche, une différence non significative a été trouvée entre les trois enfants concernant la durée du premier mot (sur les six catégories). Une anova à mesures répétées montre que cette différence n'est pas significative: F (2, 12)=1, 99 (p>.10).

La même absence de significativité a été observée entre les trois enfants concernant la durée du deuxième mot (sur les six catégories). Une anova à mesures répétées montre en effet que cette différence n'est pas significative: F(2,12) = 2, 26 (p>. 10).

4.2.2. Comparaison des pentes des patrons intonatifs - comparaison entre sujets.

Nous prenons en compte la pente de chaque courbe d'intonation lexicale (écart en quarts de ton), et non pas les valeurs brutes de fréquence fondamentale. En effet, chaque enfant a sa propre valeur de F0, distincte de celle des autres; ainsi, la fréquence du bilingue C-GH va de 250 à 450Hz, tandis que celle de l'autre bilingue, C-DY, ne dépasse pas 350Hz. Les pentes ont donc été calculées en quarts de ton.

Les résultats suggèrent que les pentes des premiers mots des énoncés à deux termes ne permettent pas de distinguer les trois sujets: une anova à mesures répétées

le montre en effet (F (2,12)=.77, n.s.). Il en va de même pour les seconds mots des énoncés à deux termes (F (2, 12)=.33, n.s.).

- comparaison entre patrons.

La comparaison des pentes associées à chacun des deux mots est effectuée en considérant le type de patron comme facteur principal. Une anova à un facteur a été menée sur le premier mot: F (2, 24)= 38, 11; p<.0001. Elle indique un effet significatif de ce facteur. L'analyse post-hoc (PLSD Fisher) montre que cet effet est principalement dû à la différence entre les Patrons 1 et 2, et entre les Patrons 2 et 3. Mais la même anova effectuée sur le second mot ne révèle aucun effet du facteur type de patron: F (2, 24)= 2,06; p>. 10.

La comparaison entre sujets montre qu'on ne trouve pas de différence concernant la pente chez ces trois enfants pour le premier et le deuxième élément de leurs énoncés à deux mots. Un tel résultat suggère que les enfants coréens de la présente étude, vivant dans une situation bilingue, ne se différencient pas d'un enfant monolingue français pour la compétence prosodique en français, bien que cette dernière soit leur langue faible. De la même manière, la comparaison entre patrons montre que les mêmes patrons prosodiques ont été maîtrisés par les deux enfants bilingues aussi bien que par l'enfant monolingue.

5. DISCUSSION ET CONCLUSION

L'un des buts de cette recherche était de montrer que des patrons prosodiques spécifiques sont effectivement associés aux catégories sémantiques des énoncés à deux mots, aussi bien par un monolingue que par des bilingues âgés de 2;2 à 3;3 vivant dans une situation bilingue. Nous avons également trouvé que ces patrons prosodiques spécifiques au français sont maîtrisés par un monolingue français ainsi que par les deux bilingues coréen-français de notre étude, bien que le français soit pour ces derniers la langue faible. Le troisième résultat indique qu'il n'y a pas de différence importante entre ces deux bilingues et le monolingue dans la réalisation des patrons prosodiques concernant le premier et le deuxième élément des énoncés à deux mots.

Ces résultats s'inscrivent dans le débat actuel concernant l'existence d'un seul système ou de deux chez les bilingues au cours de leur acquisition de deux langues. Deuchar & Quay [Deu &Qua98] ont en effet mis en évidence l'apparition de deux systèmes morpho-syntaxiques différenciés chez de jeunes bilingues anglais-espagnol (en examinant également des énoncés à deux mots). Mais elles n'ont pas traité cette question en relation avec la différenciation prosodique, qui semble émerger très tôt et rapidement (en quelques mois) chez de jeunes enfants bilingues au cours de leur développement linguistique.

En un mot, la présente étude suggère qu'à l'âge de 3 ans, après un contact de 9 mois avec le français et sous certaines conditions, la compétence prosodique en français, langue faible d'enfants bilingues coréen-français, peut être similaire à celle d'un enfant français monolingue. Bien entendu, il conviendra d'élargir notre échantillon de locuteurs bilingues pour valider ce constat et apporter de nouveaux éléments de réponse au débat en question.

BIBLIOGRAPHIE

[Blo& Lah78] Bloom. L, & Lahey (1978), *Language development and language disorders*, New York, Wiley.

[Deu& Qua98] Deuchar. M & Quay. S. (1998), One vs. two systems in early bilingual syntax: Two versions of the question, *Bilingualism: Language and Cognition,* 1 (3), 231-243.

[Han& Dom99] Han, Y. & Dommergues, J.-Y. (1999), The early acquisition of prosody in bilinguals' weaker language, *14th International Congress of Phonetic Sciences*, San Francisco, 1459-61.

[Ing81] Ingram, D. (1981), The emerging phonological system of an Italian-English bilingual child, *Journal of Italian Linguistics*, 2, 95-113.

[Par& Gen96] Paradis, J., & Genesee, F. (1996), Syntactic acquisition in bilingual children: Autonomous or interdependent?, *Studies in Second Language Acquisition*, 18, 1-25.

Contribution à la quantification du degré d'organisation des systèmes vocaliques.

K. Huet, B. Harmegnies

Université de Mons-Hainaut
Département de Communication Parlée – Place du Parc, 18 – MONS, Belgique
Tél.: +32 (0)65 37 31 44 - Fax: +32 (0)65 37 31 42
Mél : Kathy.Huet@umh.ac.be, Bernard.Harmegnies@umh.ac.be
http://www.umh.ac.be/~compa/

ABSTRACT

This paper presents an exploratory study on the vocalic differences which can be observed under different speaking styles. A single speaker was recorded on 4 situations with gradual involvement of the speaker in the communication process. A new index is introduced for the assessment of the system's degree of organization in each speaking style. The results are compared with those drawn from a procedure based on discriminant functions.

1. INTRODUCTION

Durant les deux dernières décennies, divers travaux se sont intéressés aux variations des caractéristiques acoustiques des sons de parole qu'induisent des changements dans les conditions de production. Les premières recherches en la matière se sont principalement centrées sur l'étude de la variabilité des timbres vocaliques de diverses langues d'origine germanique [Koo80] et [Nor86]. D'autres travaux, notamment conduits par l'un d'entre nous, se sont attachés aux voyelles de diverses langues d'origine romane, telles l'espagnol [Har92], le catalan [Ble93], l'italien [Har95] et le portugais [Del97]. Peu de travaux se sont par contre centrés sur le français.

La plupart de ces recherches avaient pour objectif la confirmation de l'existence d'un effet systématique de la situation de production sur le signal de parole. Elles ont, dès lors, recouru, le plus souvent à deux modalités fortement différenciées : d'une part, la conversation *spontanée*, entretenue le plus naturellement possible avec l'informant, afin d'obtenir une qualité de production proche de celle qui caractérise ordinairement les échanges verbaux banals; d'autre part, une situation classique de production de sons de parole en *laboratoire*, faisant appel tantôt à la réalisation de logatomes, tantôt à celle de mots signifiants isolés.

Il ressort de ces études que -dans le cas des productions en langues romanes, à tout le moins- la différenciation acoustique entre les segments associés à des catégories phonologiques différentes est plus tranchée en parole *de laboratoire* qu'en parole *spontanée*. Projetés dans le plan F_1/F_2, les nuages vocaliques apparaissent ainsi plus mêlés dans le cas de la parole spontanée que dans celui de la parole de laboratoire.

Diverses modalités de quantification ont été envisagées pour rendre compte de ces observations. L'idée de *centralisation* du système vocalique en parole spontanée (fortement débattue par certains auteurs: cf. [Lin63] et [Lin90]) a ainsi fait l'objet de l'élaboration d'un indice de centralisation [Har92]. Celui-ci indique, dans la plupart des langues étudiées, une tendance dominante -quoique non omniprésente- à un rapprochement acoustique du timbre de schwa, voyelle théorique définie comme le résultat de l'excitation d'un tube de section uniforme de 17.5 cm de long et habituellement considérée comme le modèle de la voyelle neutre. Par ailleurs, il a été noté que la dispersion au sein des nuages de points dans le plan formantique est, en règle générale, plus importante en parole spontanée qu'en parole de laboratoire. La conjonction de ces deux tendances aboutit à l'idée d'un système vocalique moins différencié (présentant des nuages plus rapprochés les uns des autres, mais chacun plus dispersé) en parole spontanée qu'en parole de laboratoire.

Néanmoins, bien plus que la centralisation, c'est en fait la variation du degré d'*organisation-désorganisation* du système qui apparaît comme le phénomène le plus universel au travers de l'ensemble des expériences menées à ce jour. Il semble ainsi raisonnable de considérer que tout état d'un locuteur résultant de l'ensemble des variables de la situation d'émission peut être associé à un état de plus ou moins grande organisation/désorganisation du système vocalique.

Deux conditions au moins sont requises à la mise à l'épreuve de cette conception. D'une part, il importe de diversifier les situations de production de signal de parole, afin d'étayer l'étude sur une variabilité suffisante des styles de parole. D'autre part, il convient de se doter d'une mesure apte à rendre compte avec finesse des variations de l'état du système vocalique sous l'effet des changements de style de parole. Or, la technique utilisée jusqu'à présent (simulation d'une tâche de reconnaissance automatique sur base de fonctions discriminantes

dérivées des valeurs formantiques) ne donne finalement qu'une information quantitative indirecte et assez grossière sur la variable à l'étude.

La présente communication constitue une première contribution au comblement de ces lacunes. Elle procède ainsi à l'exposition d'un sujet francophone à quatre situations contrastées de production du signal de parole et présente une alternative (l'indice κ) à la procédure basée sur l'analyse discriminante pour l'évaluation du degré d'organisation/désorganisation du système vocalique dans les quatre situations.

2. PROCÉDURE EXPÉRIMENTALE

2.1 Corpus et recueil des productions

Le corpus est constitué de productions vocales d'un locuteur francophone de Belgique issu de la région de Mons (zone d'influence picarde) et placé dans 4 situations de communication différentes.

Dans un premier temps, le locuteur participe à une conversation ordinaire avec l'expérimentateur qui lui demande de parler de différents thèmes tels que sa situation familiale, le dernier livre qu'il a lu, ce qu'il pense de tel ou tel sujet d'actualité, etc..

Cette conversation, d'une heure, environ, est entièrement enregistrée puis transcrite orthographiquement. Un échantillon aléatoire d'au moins 45 réalisations de chaque voyelle étudiée est sélectionné. Pour la présente étude, seules les voyelles délimitant la périphérie du triangle vocalique (/i/, /e/, /ɛ/, /a/, /ɺ/, /o/ et /u/) sont prises en considération. Même si l'équipartition des effectifs entre catégories vocaliques est souhaitée, il arrive dans certains cas - la probabilité d'apparition de certains phonèmes étant faible – que le nombre d'occurrences apparaissant dans la conversation n'atteigne pas 45. C'est alors la totalité des réalisations disponibles qui sont retenues.

L'échantillon de 268 réalisations enregistrées à la faveur de la première situation de communication (que nous appellerons 'parole spontanée') constitue l'ensemble des mots cibles que nous utiliserons dans les 3 autres situations de communication. Afin de neutraliser les effets contextuels, le locuteur devra reproduire, dans les autres conditions de communication, les mots ainsi extraits du corpus recueilli lors de l'entretien spontané. Chacun des mots cibles contenant les réalisations sélectionnées sera donc disponible pour analyse dans chacune des situations de communication.

Lors de la seconde mise en situation de communication, il est simplement demandé au locuteur de lire les mots contenant les phonèmes préalablement sélectionnés. Afin d'éviter les effets de liste, chaque mot est présenté en isolation sur une fiche ; les fiches sont présentées en ordre aléatoire. On obtient ainsi le corpus de 'parole de laboratoire'.

Dans un troisième temps ('production de parole sous condition de bruit'), le locuteur est équipé d'un casque diffusant dans les deux oreilles un bruit uniformément masquant. La tâche consiste également en la production des mots cibles.

Pour la dernière situation de communication ('production de parole sur requête d'un interlocuteur'), un interlocuteur éloigné, dont notre locuteur peut percevoir la voix dans son casque, lui fait répéter les mots cibles sous prétexte d'une mauvaise compréhension. Dans ce cas précis, le stock de mots cibles est noyé dans un ensemble de distracteurs.

Ces 4 sessions d'enregistrement ont été effectuées dans la chambre sourde du Laboratoire de Phonétique de l'Université de Mons-Hainaut. Les productions du sujet ont été enregistrées grâce à un microphone Neumann U87UP48 relié via un NAGRA IV S à la chaîne d'acquisition constituée d'un digitaliseur PCM Sony et d'un magnétoscope Panasonic NH75 HQ.

2.2 Analyse acoustique

Les fréquences des premier ($F_1(j)$) et second ($F_2(j)$) formants estimées au centre de chaque voyelle (j) sélectionnée pour chaque situation de communication sont évaluées au moyen de l'analyseur CSL 4300B KAY et du logiciel d'analyse KAY Multi-Speech 3700 version 2.01, à partir de spectrogrammes à large bande.

Nous disposons donc de 1072 mesures (4 situations fois 268 réalisations, ventilées comme suit : 45 [i], 37 [e], 42 [ɛ], 45 [a], 45 [ɺ], 22 [o] et 32 [u]) et nous nous intéressons ici aux nuages de points formés dans le plan biformantique par ces mesures.

Chaque mesure f en Hertz a été convertie en mels, d'après la formule

$m = 2595 \log (1+f/700)$,

proposée par Van Bergem [Van93].

3. ANALYSE DES RÉSULTATS

3.1 Analyse discriminante

De même que dans les travaux précédemment cités, nous avons appliqué à nos mesures la procédure basée sur l'analyse discriminante. Pour chaque situation de communication, les phonèmes étudiés étaient considérés comme les catégories à priori, tandis que les fréquences formantiques F_1 et F_2 de leurs réalisations, pour une situation donnée, étaient considérées comme variables discriminantes. Une fois les fonctions de discrimination obtenues, nous les avons utilisées pour simuler une tâche de reconnaissance dans chacune des situations étudiées. De cette manière, chaque point du plan F_1/F_2 représentant une réalisation d'une voyelle

était affecté à posteriori à l'une des 7 catégories phonémiques à l'étude.

Les matrices de confusion, que nous ne pouvons reproduire ici, font apparaître des taux de reconnaissance correcte respectivement égaux à 52.6% en production de parole spontanée, 73.5% en production de parole de laboratoire, 72.4% en production de parole sous condition de bruit et 75% en production de parole sur requête d'un interlocuteur.

Au vu de ces résultats, il apparaît donc que la parole spontanée est associée avec un état de désorganisation du système nettement plus prononcé que dans les 3 autres styles. Ces derniers font, quant à eux, apparaître une suite ordonnée (en ordre croissant du pourcentage de reconnaissance correcte) : production de parole sous condition de bruit - production de parole de laboratoire - production de parole sur requête d'un interlocuteur.

Les pourcentages correspondants se différencient cependant très peu (différence maximale de 2.6%).

3.2 Indice κ

Ainsi qu'on l'a souligné plus haut, les informations quantitatives recueillies au moyen de l'analyse discriminante sont assez grossières. D'une part, ce n'est qu'à la faveur d'un raisonnement indirect que l'on considère que, si la reconnaissance est mauvaise, c'est que le système est désorganisé. D'autre part, le raisonnement est de l'ordre de l'inclusion ou de l'exclusion d'un élément par rapport à un groupe, et on peut donc suspecter une sensibilité particulière de la procédure aux valeurs exceptionnelles.

En vue de pallier ces défauts, nous avons mis au point une nouvelle procédure, inspirée de l'analyse de variance. Celle-ci se base sur l'établissement d'une analogie entre d'une part, l'*écart* (en analyse de variance) d'une valeur -donnée ou moyenne- à la moyenne de référence et, d'autre part, la *distance euclidienne* (pour nos voyelles) entre un point dans le plan F_1/F_2 –voyelle ou centre de gravité d'un nuage vocalique- et le centre de gravité de référence. Le but est de quantifier, au moyen d'un indice que nous dénommerons κ, et qui s'inspire de la statistique F de Fisher-Snedecor, le rapport de la variabilité inter-classe (i.e., inter-nuage) à la variabilité intra-classe (i.e., intra-nuage).

Nous disposons, à cet effet, de 7 groupes (constitués des 7 catégories de voyelles) et pour chacun de ces groupes, nous pouvons poser :

N_k (nombre de productions disponibles pour la catégorie k , avec 0<k<8)

$N = \sum_{k=1}^{7} N_k$ (nombre total de productions, toutes catégories confondues)

d_{intra} (distance entre le point *j* du nuage k et le centre de gravité de celui-ci).

$$d_{intra}(j) = \left[\left(F_1(j) - \langle F_1(k) \rangle\right)^2 + \left(F_2(j) - \langle F_2(k) \rangle\right)^2 \right]^{1/2}$$

avec $\langle F_1(k) \rangle$ la moyenne des F_1 de tous les points formant le nuage k et $\langle F_2(k) \rangle$ la moyenne des F_2 de tous les points formant le nuage k.

d_{inter} (distance entre le centre de gravité du nuage k et celui de l'ensemble des nuages de points).

$$d_{inter}(k) = \left[\left(\langle F_1(k) \rangle - \langle F_1 \rangle\right)^2 + \left(\langle F_2(k) \rangle - \langle F_2 \rangle\right)^2 \right]^{1/2}$$

avec $\langle F_i \rangle$ la moyenne des F_1 , tous nuages confondus et $\langle F_2 \rangle$ la moyenne des F_2, tous nuages confondus.

Ces distances étant considérées comme des écarts, nous pouvons calculer des 'sommes de carrés d'écarts' (SCE) :

$$SCE_{intra} = \sum_{k=1}^{7} \sum_{j=1}^{N_k} d^2_{intra}(j)$$

$$SCE_{inter} = \sum_{k=1}^{7} \left(N_k * d^2_{inter}(k)\right)$$

avec leurs nombres de degrés de liberté respectifs :

$L_{intra} = N - k$ et $L_{inter} = k - 1$

Enfin, nous pouvons calculer l'indice κ pour chacune de nos situations de communication :

$$\Phi = \frac{CM_{inter}}{CM_{intra}}$$

où les carrés moyens (CM) sont obtenus en divisant les SCE par le nombre de degrés de liberté correspondant :

$$CM_{intra} = \frac{SCE_{intra}}{L_{intra}} \quad \text{et} \quad CM_{inter} = \frac{SCE_{inter}}{L_{inter}}$$

Appliquée à nos données, cette procédure fournit les valeurs suivantes : 3463 en production de parole spontanée, 8970 en production de parole de laboratoire 7572 en production de parole sous condition de bruit et 10792 en production de parole sur requête d'un interlocuteur.

Au vu de ces résultats, il apparaît que peut être opéré le même classement des situations de production que celui qui avait émané du traitement à base d'analyse discriminante. Ainsi que le montre la figure 1, la répartition des valeurs obtenues sur le domaine de définition de la variable est cependant plus régulière.

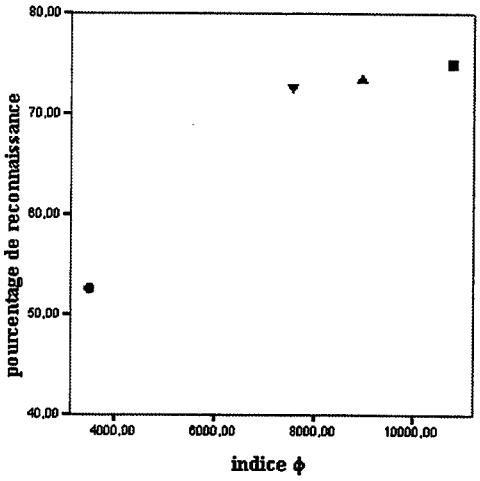


Figure 1: Taux de reconnaissance correcte (%) obtenus par l'analyse discriminante en fonction des indices κ calculés pour chacune des situations de communication envisagée : ● production de parole spontanée, ▲ production de parole de laboratoire, ▼ production de parole sous condition de bruit et ■ production de parole sur requête d'un interlocuteur.

4. DISCUSSION

Les observations effectuées à la faveur de l'expérience relatée ici confirment celles découlant des expériences antérieures, singulièrement pour les condition de *parole spontanée* et *parole de laboratoire*, dont le contraste s'apparente à ceux déjà relevés. Les pourcentages de reconnaissance correcte dérivés de la tâche de simulation de reconnaissance apparaissent en effet dans un rapport comparable à ceux déjà identifiés. Les indices κ rendent compte également du même phénomène.

En outre, les deux situations nouvelles (*production de parole sous condition de bruit* et *production de parole sur requête d'un interlocuteur*) font apparaître, pour la première, un état de désorganisation du système *intermédiaire* à ceux qui caractérisaient la parole spontanée et la parole de laboratoire et, pour la seconde, un état d'organisation *supérieur* à celui correspondant à la parole de laboratoire. Le classement des situations de production de parole qui peut ainsi en être dérivé (1. production de parole sur requête d'un interlocuteur; 2. production de parole de laboratoire; 3. production de parole sous condition de bruit; 4. production de parole spontanée) est identique, que l'on s'appuie sur les pourcentages de reconnaissance correcte dérivés de l'analyse discriminante ou sur les indices κ. Néanmoins, on note que la répartition des valeurs de κ sur le domaine de définition de la variable est plus régulière que celle des pourcentages de reconnaissance. En effet, si ces derniers montrent une nette différence entre la parole spontanée et les trois autres styles de parole, ils se singularisent par une concentration importante de ces trois valeurs autour de celle de la parole de laboratoire.

Bien que ces observations, établies à partir d'un nombre de données trop peu important ne permettent pas de tirer de conclusion définitive, elles paraissent cependant justifier la poursuite des investigations relatives à une quantification du degré d'organisation/désorganisation du système vocalique à base de techniques numériques dérivées de l'analyse de variance.

Par ailleurs, si jusqu'à présent, le signal de parole produit artificiellement sur injonction d'un expérimentateur était apparu comme la plus proche manifestation du système canonique du locuteur, on observe ici que les signaux de parole émis sur requête de clarification d'un interlocuteur éloigné s'associent à un état encore plus différencié du système.

BIBLIOGRAPHIE

[Koo80] KOOPMANS-VAN BEINUM F. (1980) "Vowel Contrast Reduction. An Acoustic and Perceptual Study of Dutch Vowels in Various Speech Conditions", Amsterdam Academische Presse.

[Nor86] NORD L. (1986), "Acoustic studies of vowel reduction in Swedish", STL-QPSR 4, pp. 19-36.

[Har92] HARMEGNIES B., POCH-OLIVE D. (1992) "A study of style-induced vowel variability: laboratory versus spontaneous speech in Spanish", Speech Communication, 11, pp. 429-437.

[Ble93] BLECUA-FALGUERAS B., POCH-OLIVE D., HARMEGNIES B. (1993), "Variaciones en la organización de la vocales del español y del catalán en function del estilo de habla", Proceedings of the International Conference of Applied Linguistics, Granada, pp. 97-107.

[Har95] HARMEGNIES B., POCH-OLIVE D. (1995), "A dynamic Approach of Vowels Systems in Italian", Proceedings of the XIIIth International Congress of Phonetics Sciences, Stockholm, 1, pp. 408-412.

[Del97] DELPLANCQ V., HARMEGNIES B. (1997), "Une modélisation à base angulaire pour l'étude de la réduction vocalique", Actes du 4ème congrès français d'acoustique, Toulouse, Teknea, pp. 389-393.

[Lin63] LINDBLOM B. (1963), "Spectographic Study of Vowel Reduction", J. Acoust. Soc. Amer. 35, pp. 1771-1781.

[Lin90] LINDBLOM B. (1990), "A sketch of the H and H theory". In A. Marchal et W. Hardcastle (eds.). Speech Production and Modelling pp. 403-440, Dordrecht, Kluwer, Academic Publishers.

[Van93] VAN BERGEM D.R. (1993), "Acoustic vowel reduction as a function of sentence accent, word stress, and word class", Speech Communication 12, pp. 1-23.

A propos de la catégorisation fonctionnelle des kinèmes co-verbaux

Jean Marc COLLETTA

Lidilem, Université Stendhal-Grenoble III
B.P. 25 - F-38040 GRENOBLE CEDEX 9
tel. + fax : 04.76.74.73.67
mél : jean-marc.colletta@u-grenoble3.fr

1 INTRODUCTION

Nous sommes actuellement engagés dans une exploration des conduites langagières enfantines orales, dans une approche multimodale et avec une perspective développementale. A cette fin, nous disposons de 7h30 d'enregistrements vidéo au cours desquels 60 enfants âgés de 6 à 12 ans ont été filmés, par groupes de trois, en conversation avec un adulte.

Les premières exploitations de ces données on fait apparaître une étroite solidarité entre la parole et le langage du corps [Col98a, Col98b, Col99]. En ce qui concerne la posturo-mimo-gestualité co-verbale, tout indique que celle-ci varie à la fois en fonction de l'âge et en fonction du type d'activité langagière en cours.

Pour vérifier l'impact de ces deux sources de variation, nous avons cherché à catégoriser les kinèmes co-verbaux (gestes, mimiques, changements de posture et de regards fonctionnellement reliés à la parole) produits par les enfants. Un classement empirique a donc été établi à partir des données exploitées, classement qui recoupe ceux établis par d'autres auteurs [Cos93, Sch84] pour la posturo-mimo-gestualité adulte. Cette classification est présentée dans le tableau 1 en annexe. Elle permet de faire la distinction entre les mouvements "autonomes", produits en l'absence de la parole et qui se substituent à elle, et les mouvements co-verbaux, "associés" à la parole.

2 UNE ÉTUDE PRÉALABLE

Cette classification construite, il restait à en tester la validité. Sur ce point, nous n'avons pu nous appuyer sur des études antérieures. En effet, si la littérature relative au non verbal fait état de nombreuses études consacrées à la perception des émotions, à l'évaluation des expressions faciales, des regards et des attitudes posturales, ou à la sémiotique de la gestualité manuelle, notamment des emblèmes ([Fey85], [Cal89], [Des89], [Poy92]), en revanche, nous n'avons pas trouvé trace de travaux visant à tester l'attribution de fonctions aux mouvements co-verbaux. Comme l'a d'ailleurs noté A. Kendon : " Does gesticulation function communicatively ? It is remarkable how few investigations there are that have tackled this question " [Ken80 : 225]...

Au cours d'une étape préliminaire, nous avons demandé à 32 sujets, étudiants en sciences du langage, de catégoriser 85 kinèmes extraits de notre corpus vidéo. Les sujets visionnaient chaque séquence où était produit un mouvement (geste, mimique ou autre) afin qu'il soit replacé dans son contexte de production, et devaient attribuer une ou plusieurs fonctions (parmi les 5 principales) à ce mouvement. Au préalable, chaque catégorie était brièvement définie et illustrée par un ou deux exemples, afin que les sujets se familiarisent avec l'utilisation de la grille de codage.

Les résultats [Col98c] se sont avérés encourageants, puisque le taux d'accord entre les catégorisations effectuées par les sujets et les catégorisations théoriques effectuées a priori atteignait 72%. Cela dit, des différences assez nettes sont apparues entre les catégories, certains types de kinèmes étant très bien reconnus, d'autres nettement moins bien. Il est également apparu que les trois principales zones fonctionnelles de la kinésie associée (référentiels, méta-discursifs et syntaxiques) entretenaient des liens très étroits qu'il serait utile d'étudier de manière plus précise.

Notons par ailleurs que lors de cette étude, et pour des raisons techniques, le nombre de catégorisations de chaque mouvement était au final peu élevé (une dizaine en moyenne). En outre, tous les sujets étaient étudiants en sciences du langage, et relativement familiers des concepts sur lesquels reposent nos catégories fonctionnelles : des étudiants novices parviendraient-ils à catégoriser les kinèmes proposés aussi aisément ?

3 LA CATÉGORISATION DES CO-VERBAUX

A l'issue de cette première investigation, se posaient donc tout à la fois la question de la procédure utilisée pour tester la pertinence de nos catégories fonctionnelles, et la question des relations entre ces catégories. Aussi avons-nous résolu de tester à nouveau la validité de notre classification.

Nous avons sélectionné 13 paires de mouvements parmi les catégories de kinèmes associés : 2 mouvements par catégorie fonctionnelle, soit 12 kinèmes référentiels, 6 kinèmes méta-discursifs, 6 kinèmes syntaxiques, et 2 kinèmes phatiques. Ces mouvements ont été donnés à catégoriser, selon le même protocole que lors du premier test, à 50 sujets : 25 étudiants "linguistes" de l'université Stendhal, et 25 étudiants "non linguistes" de l'université Joseph Fourier.

Plus systématique que la première, cette seconde étude devait permettre de vérifier les 3 hypothèses suivantes :

- H1 : certains types de kinèmes sont plus faciles à catégoriser que d'autres ;

- H2 : les mouvements plus difficiles à catégoriser sont davantage perçus comme pluri-fonctionnels par les sujets ;

- H3 : les sujets "non linguistes" obtiennent des scores de reconnaissances semblables aux sujet "linguistes", et en conséquence, la classification proposée constitue un outil fiable pour une approche quantifiée de la kinésie enfantine.

4 LES RÉSULTATS

4.1 Résultats généraux

Les résultats sont globalement conformes à ceux de la première étude, avec un taux d'accords entre les catégorisations observées et les catégorisations théoriques un peu inférieur (63%). Cela dit, on retrouve les mêmes écarts entre les scores de reconnaissance des différentes catégories fonctionnelles, comme le montrent les tableaux 2 et 3 en annexe.

A la lecture de ces tableaux, on s'aperçoit en effet que :

- les kinèmes interactifs (signaux phatiques) sont parfaitement reconnus, puisque le score de reconnaissance atteint 97% ;

- les kinèmes méta-discursifs sont également bien reconnus, à 76% ;

- les kinèmes référentiels le sont un peu moins bien, à 60% (ils étaient mieux reconnus lors de la première étude puisque le score de reconnaissance était de 75%) ;

- quant aux kinèmes syntaxiques, ils sont à nouveau mal reconnus dans cette fonctionnalité, puisque le score de reconnaissance n'atteint que 47% (mais il n'atteignait que 34% lors de la première étude).

4.2 Les fonctions attribuées aux différents co-verbaux

Notre hypothèse H1 postule qu'il est plus facile d'attribuer une fonction à certains types de kinèmes qu'à d'autres. L'examen des scores de reconnaissance dans le tableau 3 suffit à lui seul à valider cette hypothèse : les mouvements phatiques et méta-discursifs sont aisément reconnus dans leur fonctionnalité respective, alors qu'à l'inverse les mouvements que nous avons catégorisés comme "syntaxiques" se voient très souvent attribués d'autres fonctions. Quant aux mouvements référentiels, s'ils sont assez bien reconnus dans l'ensemble comme participant à la désignation et à la construction de la référence, ce n'est toutefois pas systématique.

A ce stade, il est nécessaire d'examiner plus en détail les scores de reconnaissance pour chaque sous-catégorie fonctionnelle.

Parmi les kinèmes méta-discursifs (tableau 4 en annexe), les mouvements redondants par rapport à l'acte de parole (co-actifs) ou connotant l'acte de parole (connotatifs) sont très bien reconnus. En revanche, les mouvements à valeur d'emphase, qui viennent accentuer la valeur illocutoire de l'acte de parole (amplifieurs) sont moins bien reconnus dans cette fonctionnalité. Ils sont parfois confondus avec les kinèmes syntaxiques (en vertu de leur proximité avec les intensifs qui soulignent une syllabe, un mot ou un segment de l'énoncé) et les kinèmes interactifs, peut-être parce que les mimiques faciales qui composent ces kinèmes, en vertu de la forte expressivité qui les caractérise, sont perçues comme ayant une valeur phatique ou régulatrice.

Les scores de reconnaissance des kinèmes référentiels (tableau 5 en annexe) varient considérablement selon les types de kinèmes :

- les mouvements ayant pour fonction d'illustrer un référent (illustratifs), de le désigner par pointage (déictiques), ou de le mimer (mimétiques) sont très nettement reconnus dans cette fonctionnalité sémiotique ;

- ceux qui construisent dans l'espace l'univers référentiel (locatifs) ou qui pointent des objets virtuels localisés auparavant (anaphoriques) ne le sont qu'une fois sur deux ; dans les catégorisations des sujets, ils paraissent proches des kinèmes méta-discursifs (peut-être parce que les sujets ont catégorisé en même temps les expressions faciales des enfants performant les mouvements), et surtout, des kinèmes syntaxiques ;

- enfin, les mouvements qui représentent des concepts abstraits (figuratifs) sont très mal reconnus dans cette fonctionnalité et sont très souvent confondus avec les kinèmes syntaxiques (6 fois sur 10). Cela tient à la fois au fait que ces mouvements présentent un degré d'iconicité plus faible ou entretiennent avec leurs référents des liens iconiques plus indirects (métaphoriques), et à leur proximité avec les gestes de scansion et d'accentuation de la parole (les "bâtons"), puisque comme eux, ils sont très étroitement associés au flux parolier (Cosnier [Cos93] range d'ailleurs les deux types de gestes dans une catégorie unique : les "paraverbaux").

Les scores de reconnaissance des kinèmes syntaxiques (tableau 6 en annexe) varient eux aussi considérablement :

- les mouvements qui scandent la parole (rythmiques) sont très bien reconnus ;

- en revanche, les sujets accordent très souvent une fonction méta-discursive aux mouvements qui accentuent une unité (intensifs), or ces mouvements peuvent être confondus avec les amplifieurs ;

- et les changements de posture à valeur démarcative sont analysés comme ayant soit une fonction interactive (les sujets leur attribuant une valeur phatique), soit comme ayant une fonction méta-discursive.

En conclusion sur ces catégories fonctionnelles, il apparaît donc que :

- certains types de kinèmes sont aisément reconnus : les déictiques ainsi que les gestes à fort degré d'iconicité, qui participent à la désignation et à la représentation des référents ; les mimiques qui accompagnent la réalisation des actes de parole ou qui connotent le discours, les mouvements phatiques, ainsi que la gestualité rythmique étroitement associée à la parole ;

- d'autres sont moins bien reconnus et paraissent plus problématiques, qu'il s'agisse des mimiques permettant

d'amplifier la force des actes de parole, ou des gestes à valeur locative et anaphorique ;

- d'autres, enfin, sont très mal reconnus, et en tous les cas perçus comme ayant d'autres fonctions : il s'agit des gestes figuratifs, des mouvements accentuant une unité linguistique, et des posturèmes à valeur démarcative.

La figure 1 en annexe synthétise ces observations et représente les catégorisations réalisées par les sujets, en pourcentage, sur les 4 axes correspondant à nos 4 catégories principales de co-verbaux.

4.3 *Difficulté de catégorisation et pluri-fonctionnalité des mouvements*

L'hypothèse H2 postule que plus les mouvements sont difficiles à catégoriser, plus ils sont perçus comme pluri-fonctionnels par les sujets, ceux-ci ayant le choix d'attribuer plusieurs fonctions à un mouvement.

Globalement, on constate effectivement que le nombre de catégorisations effectué par les sujets est plus important pour les kinèmes dont la fonction est la moins aisément perçue (plus de 60 catégorisations en moyenne, contre moins de 55 catégorisations en moyenne pour les kinèmes aisément reconnus). Mais le calcul du Khi 2 concernant ces répartitions n'est pas significatif.

Pourtant, comme le montre la figure 1 en annexe, les mouvements les moins bien reconnus sont bien ceux pour lesquels les sujets hésitent entre plusieurs fonctions. Aussi, si la pluri-fonctionnalité de ces mouvements n'apparaît pas comme une décision "intra-sujet", elle n'en apparaît pas moins dans les décisions de catégorisations inter-sujets, certains préférant par exemple attribuer une fonction méta-discursive à un posturème démarcatif, d'autres une fonction phatique, d'autres encore une fonction syntaxique.

L'hypothèse H2 peut donc elle aussi être validée, et il y a bien une relation entre le "degré de transparence fonctionnelle" d'un kinème co-verbal et la difficulté à attribuer une fonction définie à ce kinème. Sur la figure 1, les 3 zones concentriques symbolisent cette relation, que nous théorisons de la manière suivante :

- la zone blanche à l'extérieur représente la zone communicative de la kinésie co-verbale : les kinèmes localisés dans cette zone remplissent une fonction (sémiotique, méta-discursive ou interactive) définie et aisément perceptible par l'interlocuteur ;

- la zone foncée au centre représente la zone énonciative-expressive de la kinésie co-verbale : les kinèmes localisés dans cette zone ne remplissent aucune fonction définie, sont difficiles à catégoriser et jouent probablement un rôle plus important pour le locuteur (dans le processus de production de la parole) que pour l'interlocuteur ;

- la zone grise intermédiaire rassemble des kinèmes perçus comme ayant une fonction dominante, mais également comme étant bi- ou tri-fonctionnels, leur fonction communicative n'étant pas identique pour tout le monde.

Quelques explications. Lorsque nous avons élaboré notre classification fonctionnelle des mouvements co-verbaux, nous nous sommes placés d'emblée dans la perspective sémasiologique de la réception-interprétation des signaux corporels, ce qui nous a amené à privilégier les fonctions communicatives (interactivité, méta-discours, sémiotique référentielle) de la kinésie, tout en laissant au second plan la perspective onomasiologique de la production.

Or, comme le signalent de nombreux auteurs parmi lesquels Cosnier et Brossard [Cos84] ou D. McNeill [McN92], les mouvements corporels sont d'abord et avant tout nécessaires au locuteur, dans la mesure où l'activité kinésique joue un rôle de facilitation du processus énonciatif. Le phénomène d'auto-synchronisation mis à jour par Condon et Ogston [Con84] en rend bien compte, de même que les études qui aboutissent à classer les co-verbaux en deux catégories : les "motor movements", intimement liés au processus de production de la parole, et les "symbolic movements", davantage orientés vers la sémiose et la communication [Had92].

Les difficultés rencontrées par nos sujets en matière de catégorisation des kinèmes intensifs, démarcatifs et figuratifs s'expliquent dès lors par le fait que ces types de mouvements, nécessaires à la production de la parole mais sans doute moins pertinents pour l'interlocuteur, ne sont finalement pas à leur place dans notre grille de codage, conçue pour attribuer des fonctions communicatives aux mouvements. La catégorie "syntaxique" apparaît après coup comme bien mal nommée, et comme regroupant en réalité tous les mouvements produits en étroite synergie avec le flux parolier.

Quoi qu'il en soit, le degré de transparence fonctionnelle des mouvements devient un paramètre clé de notre étude, qu'il est nécessaire de prendre en compte dans l'évaluation de la pertinence de notre classification.

4.4 *Pertinence de la classification proposée*

L'examen de l'hypothèse H3 a pour but de vérifier si des sujets "novices", non familiers des concepts sur lesquels reposent nos catégories fonctionnelles (sémiotique référentielle, méta-discours, démarcation syntaxique, interactivité, etc.) parviennent à catégoriser les kinèmes proposés aussi aisément que des sujets experts.

A cette fin, nous avons comparé les scores de reconnaissance au sein de nos deux groupes de sujets : les 25 étudiants linguistes et les 25 étudiants non linguistes. Or le Khi 2 n'est pas significatif lorsqu'on compare le score général de reconnaissance dans les 2 répartitions, même si les non linguistes obtiennent un score de reconnaissance un peu plus important (67.5%) que les linguistes (59%). Les comparaisons pour chaque grande catégorie fonctionnelle (référentiels, méta-discursifs, syntaxiques et interactifs) livrent également des scores de reconnaissance proches (les scores des non linguistes étant toujours légèrement supérieurs à ceux des linguistes), et tous les khi 2 sont non significatifs, sauf celui qui résulte de la comparaison des scores de reconnaissance des méta-

discursifs (khi 2 significatif à .01), ceux-ci étant nettement mieux reconnus par les non linguistes !

Autrement dit, les sujets novices (non familiers des concepts sous-jacents à notre classement fonctionnel) reconnaissent aussi bien, voire mieux que les sujets experts, les mouvements co-verbaux qui leur sont donnés à catégoriser. Au vu de ces résultats, on peut donc admettre la pertinence de notre classification.

Le meilleur score obtenu par les sujets novices a de quoi étonner : on pourrait croire que des étudiants en sciences du langage seraient mieux à même de classer des mouvements en fonction de leur rôle par rapport à la parole que des étudiants non linguistes. Or on observe le contraire, et tout se passe comme si les experts, connaissant la complexité de la parole et les multiples dimensions des conduites langagières, se trouvaient embarrassés face à cette tâche de catégorisation, tandis que les sujets novices classent les mouvements sans état d'âme et sans se poser de question, s'en remettant peut-être davantage à leur propre expérience de la communication multimodale.

5 CONCLUSION

Nous cherchions à tester la pertinence de notre classification fonctionnelle de la gestualité co-verbale enfantine. La présente étude, venant confirmer pour l'essentiel les résultats de l'étude préliminaire, montre que cette classification est opératoire, à condition toutefois de prendre en compte le degré de transparence fonctionnelle des kinèmes.

Les différences dans les scores de reconnaissance des différents types de kinèmes amènent à penser que les mouvements co-verbaux dont la fonction est aisément identifiable sont effectivement orientés vers le processus communicatif et la sémiose, tandis que les autres ont une orientation fonctionnelle opposée, sont étroitement associés aux processus de production de la parole, et sont en conséquence moins pertinents pour l'interlocuteur que pour le locuteur lui-même.

Des prolongements à cette étude sont dores et déjà prévus, mais l'étude des catégorisations des mouvements accompagnant la parole apparaît d'emblée comme une piste intéressante, jusque là négligée par les spécialistes du non verbal et de la communication multimodale.

BIBLIOGRAPHIE

[Cal89] Calbris G., Porcher L. (1989) Geste et communication. Credif-Hatier.

[Col98a] Colletta J.-M. (1998) "Quand les enfants racontent leur expérience avec les mots, la voix et le corps", communication au colloque Histoires de vie et dynamiques langagières, Rennes, sept.98, actes à paraître.

[Col98b] Colletta J.-M. (1998) "Les conduites narratives chez l'enfant : approche étho-linguistique et développementale", communication au 6th International Pragmatics Conference, Reims, juillet 98, à paraître dans Lidil n°22.

[Col98c] Colletta J.-M. (1998) "Catégorisation fonctionnelle des kinèmes. Etude autour d'un outil d'analyse", communication affichée au colloque ORAGE'98, Besançon, déc.1998.

[Col99] Colletta J.-M. (1999) "Quand les enfants argumentent avec les mots, la voix et le corps", communication au colloque Les relations inter-sémiotiques, Lyon, déc.99, actes à paraître.

[Con84] Condon W.S. (1984) "Une analyse de l'organisation comportementale", in Cosnier, Brossard, La communication non verbale. Delachaux et Niestlé, pp.31-70.

[Cos82] Cosnier J. (1982) "Communications et langages gestuels", in Cosnier, Coulon, Berrendonner, Orecchioni, Les voies du langage. Dunod, pp.255-304.

[Cos93] Cosnier J. (1993) "Etude de la mimo-gestualité", in R. Pléty (dir.), Ethologie des communications humaines. P.U.L., pp.103-115.

[Cos84] Cosnier J., Brossard A. (1984) "Communication non verbale : co-texte ou contexte ?", in Cosnier, Brossard, La communication non verbale. Delachaux et Niestlé, pp.1-29.

[Des89] Descamps M.-A. (1989) Le langage du corps et la communication corporelle. P.U.F.

[Fey85] Feyereisen P., de Lannoy J.-D. (1985) Psychologie du geste. Pierre Mardaga.

[Had92] Hadar U. (1992) "The dissociation between motor and symbolic movements in coverbal behavior", in F. Poyatos (ed.) Advances in nonverbal communication. John Benjamins Publishing Company, pp.113-123.

[Ken80] Kendon A. (1980) "Gesticulation and speech : two aspects of the process of utterance", in M.R. Key (ed.) The relationship of verbal and nonverbal communication. Mouton, pp. 207-227.

[McN92] McNeill D. (1992) Hand and mind. What gestures reveal about thought. The University of Chicago Press.

[Poy92] Poyatos F. (ed.) (1992) Advances in nonverbal communication. John Benjamins Publishing Company.

[Sch84] Scherer K.R. (1984) "Les fonctions des signes non verbaux dans la conversation", in Cosnier, Brossard, La communication non verbale. Delachaux et Niestlé, pp.71-100.

Les réalisations prosodiques de la focalisation en coréen spontané

Moon-Kyou PARK

Université de Paris 7-UFR de Linguistique
Tél.: 01 46 05 54 19 – Mél : parkmoon@imaginet.fr

ABSTRACT

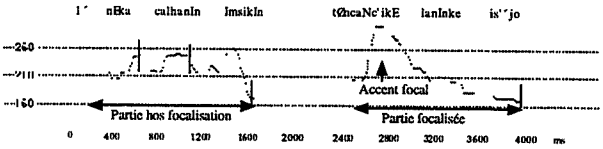
In this paper, we report the results of a preliminary study of the effects of focus on the phonetic markers in spontaneous Seoul dialect of Korean. We find that focus is phonetically marked by widening of the f0 range, slowing of the articulation rate variation and strengthening of intensity. This effect of focus is strongly marked on the focus word. On the other hand, we have not observed reduction of the f0 range, neither shortening of duration, neither weakening of intensity on the constituents preceding and following the focused word.

1. INTRODUCTION

Cette étude a pour objectif d'analyser l'influence de la focalisation sur des paramètres prosodiques tels que configurations mélodiques, profils temporels et variations d'intensité du coréen standard. Nous examinerons non seulement les séquences focalisées, mais également les séquences situées en position préfocale et postfocale.

La focalisation se définit, dans cette étude, comme la procédure de mise en relief par des facteurs prosodiques, souvent marquée par une forte montée tonale sur la syllabe non finale dans un mot (Fig.1). En effet, dans la parole spontanée, la prosodie joue un rôle important de générateur de structures énonciatives, afin de réaliser le message conformément à l'intention du locuteur. Au plan discursif, la procédure de mise en relief est considérée comme « la marque caractéristique du discours argumentatif, polémique destiné à imposer le message en faisant ressortir certains éléments du référent » [Cal87].

Figure 1 : Configurations tonales d'énoncé coréen: partie focalisée vs. partie hors focalisation.



De nombreux travaux ont été proposés pour rendre compte de la réalisation concrète de substances prosodiques sur les séquences focalisées ainsi que sur les séquences adjacentes. Ainsi, en anglais, la durée et la f0 d'un mot focalisé sont significativement plus proéminentes que d'un mot produit dans une phrase non focalisée. Les séquences postfocales manifestent, en revanche, un contour de f0 quasi plat et une durée raccourcie [Coo85]. En grec, [Bal99] ont observé des configurations tonales similaires, mais des profils temporels différents de ceux de l'anglais : la focalisation allonge la durée sur toute la phrase. En français, on observe une expansion tonale sur le mot focalisé, et sur des mots postfocaux, un contour réduit et quasi plat [Tou87] [Di99]. En ce qui concerne le profil temporel, la durée des mots porteurs d'une focalisation est plus longue que celle des mots non focalisés. La durée des mots en positions pré- et post-focales est en revanche abrégée par rapport à des mots neutres [Tou87].

En coréen, les chercheurs sont partagés quant à la manière dont s'opère la focalisation comme réalisation prosodique. Certains suggèrent que le facteur prosodique déterminant la focalisation est principalement l'allongement de la durée, non seulement sur le mot focalisé, mais aussi sur les mots adjacents [Oh98]. D'autres [Chu97] [Jun98] proposent la durée et la f0 comme des substances marquées par la focalisation. Cependant, les avis sont partagés pour le paramètre de la durée, même entre les deux chercheurs. Ainsi, à partir d'un corpus de phrases lues, [Chu97] observent un allongement systématique de la syllabe finale du mot focalisé, tandis que, selon [Jun98], l'allongement porte sur la syllabe initiale du mot focalisé. Cependant cet allongement n'apparaît pas toujours sur le mot focalisé (allongement non systématique) ; par contre, ce phénomène d'allongement réduit la durée des mots pré- et post-focaux. S'agissant de la f0, ces chercheurs observent une expansion tonale ample sur le mot focalisé et une réduction sur les mots adjacents.

De notre côté, il nous paraît intéressant d'engager notre recherche sur la focalisation dans la parole spontanée, car la parole spontanée permet d'éviter des artefacts liés aux phrases de laboratoire qui ne rendent pas compte du rôle réel de la prosodie dans les actes de communication.

Les principales questions auxquelles nous espérons pouvoir apporter une réponse au cours de cette étude sont :

1.De quelle manière s'effectue, dans la parole spontanée, la structuration temporelle et tonale des séquences focalisées et des séquences en position pré- et post-focale ? Trouve-t-on les mêmes résultats que ceux obtenus pour la parole lue par les chercheurs cités précédemment ?

2.La variation de l'intensité serait-elle un facteur important de la focalisation en coréen ?

2. PROCÉDURE D'ANALYSE

2.1 Corpus

Le corpus est constitué d'une interview réalisée par l'auteur lui-même, d'environ 10 minutes, portant sur une recette de

cuisine. Cette interview se divise en deux parties, avec des « grands tours » de parole et des « petits tours ». Afin d'obtenir les grands tours de parole, l'interlocuteur laisse le locuteur modèle (LM) achever son propos avant de lui poser une nouvelle question. L'analyse est faite essentiellement sur les « grands tours » qu'on peut considérer comme une partie monologale.

Les « grands tours » de parole ont servi à l'élaboration d'un texte à lire. La transcription a été réalisée orthographiquement.

2.2. *Sujet et enregistrement*

Nous avons enregistré un dialogue entre LM et l'auteur, en chambre calme. LM, agée de trente ans, est étudiante en Sciences de l'éducation. Née et élevée à Séoul, LM ne présente aucun accent régional. LM a été invitée à lire son texte original transcrit, en chambre sourde. La lecture a été réalisée à haute voix, avec un débit normal et une intonation « neutre », c'est-à-dire sans mettre des emphases particuliers.

Les corpus lus et spontanés ont été enregistrés sur un magnétophone portable (DAT TCD-DT de Sony) et un microphone de cravate EM166.

2.3. *Analyse et mesures*

Pour identifier les mots focalisés, nous n'avons pas pris une approche uniquement acoustique, décrivant la montée de la f0 souvent accompagnée de renforcement d'intensité, car cela représente un risque, pour nous, de confusion entre les phénomènes de focalisation et ceux d'accents purement rythmiques, appelé souvent « accent principal » ou « accent secondaire (ou ictus mélodique) ».

Pour séparer les phénomènes de focalisation des phénomènes d'accents purement rythmiques et démarcatifs, nous avons procédé à des contre-vérifications comme celles que [Rig70] a effectué dans son étude contrastive de l'accent français et de l'accent tchèque. Ces contre-vérifications qualifiées, par l'auteur, comme le seul moyen de distinguer l'accent focal de l'accent non focal, consistent à demander à des auditeurs natifs leur impression de la présence ou l'absence d'une intention particulière d'insistance sur des mots.

L'analyse acoustique a été réalisée à l'aide du logiciel Signalyze3.12. et de ses trois représentations graphiques, le spectrogramme, la courbe de f0 et la courbe d'intensité. Les mesures ont été effectuées manuellement sur un enregistrement de trois minutes pour chaque type de parole. Les durée syllabiques ont été mesurées à partir de spectrogrammes. Les f0 ont été mesurées sur la partie centrale de voyelles, et les dB, sur les pics d'intensité de voyelles. La fluctuation de l'intensité relative a été obtenue par une normalisation de valeurs réduites z.

3. RÉSULTATS ET DISCUSSION

3.1. *Configurations tonales*

La représentation tonale de parole lue que[Chu97] [Jun98] avaient attribuée à la séquence focalisée et aux séquences adjacentes est partiellement confirmée par nos observations.

En position focale, le mot porteur de la focalisation se caractérise par une configuration tonale fort différente, en comparaison avec son homologue de parole lue : l'expansion tonale est relativement ample. L'intervalle fréquentiel du mot focalisé et du mot neutre est de 59 Hz vs. 32 Hz. Le test-t montre que cette différence est très significative : p (64)= <,0001. L'amplification enregistrée sur le mot focalisé est dû, principalement, à une relative augmentation de la fréquence du maximum (264 Hz contre 225 Hz: p (64)= <,0001). Son minimum focal est légèrement plus haut que son homologue produit dans la parole lue. La différence entre ces deux minima n'est pourtant pas significative. Ces résultats rejoignent [Par94] qui avait constaté que la focalisation n'influence pas le minimum du mot focalisé. La même tendance a été également observée en français lu[Tou87].

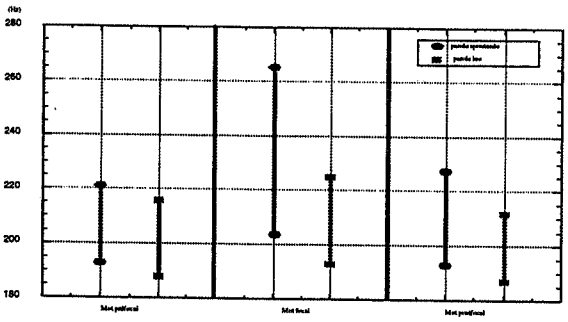


Figure 2 : Expansions tonales (moyennes des maxima et des minima de f0) en parole spontanée et lue sur les mots préfocaux, focaux et postfocaux.

Table 1 : Moyennes (avec écart-type) des maxima et des minima de f0 (les valeurs soulignées indiquent la différence significative).

Positions	préfocale		focale		postfocale	
	spon	Lu	spon	lu	spon	lu
Minimum	194 (20)	188 (12)	205 (28)	193 (21)	194 (18)	187 (17)
Maximum	220 (24)	216 (15)	<u>264</u> (27)	225 (20)	<u>226</u> (25)	212 (15)
Intervalle	26 (20)	27 (12)	<u>59</u> (31)	32 (21)	31 (24)	26 (14)

En position préfocale, il n'y a pas de différence d'intervalle fréquentiel de maximum et de minimum entre parole spontanée et parole lue. Cependant, le registre fréquentiel du maximum et du minimum de parole spontanée est légèrement plus élevé que celui de parole lue.

Sur le mot situé en position postfocale, on observe une configuration tonale similaire à celle du mot préfocal : l'intervalle fréquentiel est quasi identique entre la parole spontanée et la parole lue. Une différence réside pourtant sur le maximum de la F0 de parole spontanée, qui est significativement plus élevé que son homologue de parole lue. Cette tendance a été également observée par [Chu97] .

Si l'on compare la configuration tonale des mots selon les trois positions possibles dans la parole spontanée, le mot focalisé est plus saillant que les mots adjacents, avec une expansion tonale très importante, due principalement à son maximum de f0. Selon l'analyse de variance, ce maximum est significativement plus important que les maxima de f0 appartenant aux mots adjacents ($F(2,78)=13,171$; $p<,0001$).

Par la focalisation, le registre tonal de la locutrice est donc conjointement obtenu par une élévation des maxima et des minima. Sur ce point, nos résultats rejoignent ceux de [Bru90] qui avait souligné l'importance de la variation globale de registre mélodique en suédois spontané.

D'autre part, cette importante élévation sous l'effet de la focalisation pourrait constituer une des causes primordiales de la rupture de la ligne de déclinaison, ce qui caractérise également l'intonation de parole spontanée. Ainsi, [Mor99] a suggéré cette possibilité de l'effet de focalisation dans son étude de comparaison de parole spontanée et de parole lue.

3.2. Profils temporels

Le profil temporel est représenté ici par un débit d'articulation. Cette procédure s'est avérée nécessaire dans la mesure où les valeurs absolues en millisecondes (ms) ne facilitent pas la comparaison des durées des groupes de différente longueur.

L'analyse des différentes catégories de profils (préfocal, focal et postfocal) est essentiellement basée sur les profils de parole lue.

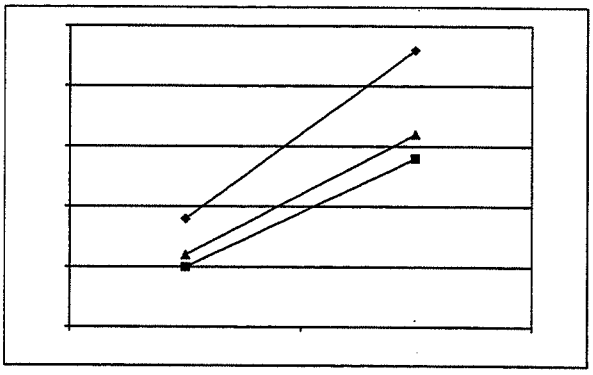


Figure 3 : Profil temporel : débit d'articulation (nombre de syllabe par seconde) sur les mots situés en position préfocale, focale et postfocale.

Les valeurs prises par le débit d'articulation varient en fonction de la position des mots dans le corpus spontané. L'analyse de variance montre que globalement la position a un effet significatif sur le débit d'articulation ($F(2,61)=5,004$; $p•0,0054$). Le mot focalisé est caractérisé par un ralentissement par rapport aux mots adjacents. Cependant, le débit du mot postfocal n'est pas significativement différent de celui du mot focal. Le mot préfocal est produit par un débit plus accéléré que le mot focalisé et le mot postfocal.

Cette tendance peut être dans une certaine mesure reliée aux observations de [Dan97] qui observe que le débit d'articulation est corrélé à un degré d'information. En effet, ces auteurs remarquent que les locuteurs renforcent l'apport de nouvelles informations, appelé souvent « rhème », par la diminution de débit d'articulation. En revanche, une information prévisible et donnée est associée à une accélération de débit.

3.3. Intensité

La comparaison entre l'intensité des mots à des positions différentes de parole spontanée et l'intensité de ces mêmes mots dans le profil neutre montre que les différences ne sont pas significatives.

Table 2 : Moyennes des maxima et des minima d'intensité en valeurs réduites z.

Positions	préfocale		focale		postfocale	
	lu	spon	lu	spon	lu	spon
Minimum	-0,5	-0,8	-0,5	-0,3	-0,6	-0,6
Maximum	0,7	0,4	0,9	1,1	0,8	0,6
Intervalle	1,3	1,2	1,4	1,4	1,4	1,2

Néanmoins, si l'on compare l'intensité de séquences focalisées et celle de séquences adjacentes dans le corpus spontané, les effets d'intensité de la focalisation semblent être distribués sur la séquence porteuse de la focalisation. L'analyse de variance montre que le maximum de l'intensité enregistré sur le mot focalisé est significativement plus important que le maximum de mots adjacents. Cependant l'écart de maximum et de minimum d'intensité du mot focalisé ne se révèle pas différent de l'écart des mots adjacents. Cet ajustement inattendu provient du fait que la focalisation relève non seulement du maximum, mais aussi du minimum de son mot focal, ce qui nous rappelle la configuration tonale du mot focal.

Les effets d'intensité de focalisation semblent être également distribués en grande partie sur les syllabes suceptibles de porter l'accent focal, marquées par une montée mélodique abrupte [Koo86]. En effet, cette distribution représente 92,5% de cas dans notre corpus. Il paraît donc évident que les effets de f0 et ceux de l'intensité sont fortement corrélés dans la focalisation.

4. CONCLUSION

Dans cette étude, nous avons montré qu'en coréen spontané, la focalisation est associée aux paramètres prosodiques tels que l'élargissement du registre tonal, le ralentissement de débit d'articulation et le renforcement de l'intensité. Cet effet de focalisation est fortement marqué sur le mot focalisé. Par contre, nous n'avons pas observé, d'une manière régulière et systématique, ni de réduction tonale, ni d'abrègement temporel, ni d'affaiblissement d'intensité sur les mots adjacents. Une analyse acoustique et perceptive plus détaillée du niveau des syllabes, et éventuellement d'une unité plus grande tel que le groupe prosodique, serait nécessaire afin de mieux appréhender les systèmes prosodiques dans l'organisation énonciative.

BIBLIOGRAPHIE

[Bal99] Baltazani M. & Jun S. (1999), " Focus and topic intonation in Greek ", ICPhS99, pp.1305-1308.

[Bru90] Bruce G. & Touati P. (1990), " On the analysis of prosody in spontaneous dialogue", Working papers, 36, Lund University, pp.37-55.

[Cal87] Callamand M. (1987) " Analyse des marques prosodiques du discours ", Etudes de Linguistique Appliquée, N°66, pp.49-70.

[Chu97] Chung S.J. & Kenstowicz M. (1997), " Focus expression in Seoul Korean", Havard Studies in Korean Linguistics, pp.93-105.

[Coo85] Cooper W. & Mueller P. (1985), " Acoustical aspects of contrastive stress in question-answer contexts ", JASA, Vol77(6),pp.2142-2156.

[Dan97] Dankovicova J. (1997), " The domain of articulation rate variation in Czech", Journal of Phonetics, Vol.22 pp.287-312.

[Di99] Di Cristo A. & Jankowski L. (1999), " Prosodic organisation and phrasing after focus in French", ICPhS99, pp.1565-1568.

[Jun98] Jun S.A. & Lee H.J. (1998), " Phonetic and phonological markers of contrastive focus in Korean", ICSLP98, pp.2323-2326.

[Koo86] Koo H.S. (1986), An experimental acoustic study of the phonetics of intonation in standard Korean, Hanshin Pub.

[Mor99] Moraes J. (1999), " F0 declination in Brazilian portuguese in read and spontaneous speech", ICPhS99, pp.2323-2326.

[Oh98] Oh M. (1998), A Korean prosodicstructure and focus, Yeojoo Institute of Technology.

[Par94] Park M.K. (1994), Structure prosodique du coréen et du français , mémoire de DEA,Université de Paris 7.

[Rig70] Rigault A. (1970), " L'accent dans deux langues à accent fixe", Studia Phonetica, Didier, pp.1-12.

[Tou87] Touati P. (1987), Structures prosodiques du suédois et du français, Lund University Press.

Etude comparative de la palatalisation et des palatales (français et coréen)

Hyeon-Zoo KIM

Department of French Language and Literature, Dankook University
29, Anseo-dong Cheonan Choong-nam, South Korea
Tél. & Fax.: ++82 (0)343 444-5161
Mél: hyeonzoo@kornet.net

ABSTRACT

In a palatalizing context (+i, j), Korean and French plosives were analyzed according to three variables : VOT (Voice Onset Time), duration of consonant and duration of preceding vowel.
Moreover the [a] vowel influences least surrounding consonants in contrast with the [i] vowel which tends to palatalize consonants, in French especially.
In French, there are no unvoiced consonants but palatalization occurs ; whereas in Korean, there are palatals, but no palatalization occurs.
Results point to systematic differences in Korean and French as to the amplitude and palatalization degree. Why do elements in a similar context behave differently?
An explanation involves the concept of phonetic context and its consequences ; it involves as well the crucial concept of system which must be combined with the concept of "phonological constraint."

1. INTRODUCTION

Le plupart des adultes qui apprennent une langue étrangère parlent avec un "accent" qui provient partiellement du contact entre la phonologie et la phonétique propre à leur langue maternelle (L1) et celle qui caractérisent la langue étrangère dont ils font l'apprentissage. Entre la langue maternelle (L1) qu'on apprend au début de l'enfance et une langue secondaire (L2) qu'on apprend plus tard dans la vie, une influence s'exerce réciproquement. Ceci a été démontré aussi bien dans le domaine sémantique (Obl78 ; Mac89). que dans le domaine phonologique (Fle91). Par contre, rien ne prouve encore qu'il en soit de même dans le domaine phonétique, du moins pour les individus qui apprennent deux langues au début de l'enfance.

Le système consonantique du coréen comporte trois séries d'occlusives sourdes /p, t, c, k, p^h, t^h, c^h, k^h, p', t', c', k'/. Il est admis qu'en français les deux séries d'occlusives s'opposent par la sonorité (sourdes/sonores) ou par la tension (fortes/faibles) /p, t, k, b, d, g/ (Kea84).

Nous voulions donc savoir si la durée, facteur acoustique présent en permanence, était un corrélat nécessaire voire suffisant de la distinction entre les occlusives. Et les Coréens, en produisant des phonèmes français, utilisent-ils leur propre système? ou a contrario imitent-ils le système français? Y a-t-il des consonnes palatalisées chez les Coréens parlant français?

L'analyse des matériaux de la présente étude nous a permis, d'une part, d'apporter quelques compléments aux résultats d'autres études et, d'autre part, d'en d'obtenir de nouveaux particulièrement intéressants en ce qui concerne le VOT. Le VOT est un indice universel, mais il n'est pas à considérer dans sa singularité. Selon les références (Zim58 ; Lae92), il faut le prendre en compte dans le cadre de son contexte, il doit être mis en rapport étroit avec les données de la durée de la tenue et de la durée totale.

Le VOT correspond à ce que nous appelons habituellement l'explosion dans la mesure où nous distinguons trois phases dans les occlusives : implosion, tenue, explosion. Premièrement, l'implosion qui est la mise en place des organes. Deuxièmement, la tenue, avec la fin de la tenue, rupture de l'occlusion et début du VOT. Troisièmement, l'explosion, en soulignant toutefois qu'en ce qui concerne les aspirées, il faudrait distinguer ce qui correspond à l'explosion et ce qui correspond à l'aspiration (Kim94).

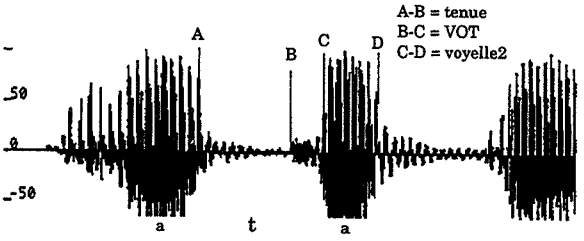


Figure 1. Tracé d'Audiomedia

2. PRELIMINAIRES

2.1. Méthode

Dans cet article, nous consacrerons la plus grande partie de notre étude aux exploitations des mesures effectuées à partir des documents oscillographiques mesurés avec le système "Audiomedia" sur un ordinateur Macintosh. Deux méthode complémentaires, les mingogrammes à 4 lignes et l'Audiomedia, ont été utilisées pour obtenir des renseignements sur :

- la sonorité : 4 lignes
- le VOT : Audiomédia
- la hauteur d'explosion : 4 lignes (plus significatif)
- la durée vocalique : Audiomédia et 4 lignes
- la durée consonantique : Audiomédia et 4 lignes

L'Audiomedia enregistre directement sur le disque dur, à partir duquel se fait également la lecture. Les processeurs de la carte "audiomedia" améliorent énormément les

capacités audios de l'ordinateur, tout en lui permettant de gérer les tâches exigeantes en temps de calcul.

2.2. *Choix du corpus et des sujets*

Dans notre travail consistant en l'étude descriptive du VOT des consonnes occlusives nous nous attacherons surtout à placer via des phrases et expressions, les articulations consonantiques, dans les principales positions où elles peuvent figurer en coréen et en français par rapport à l'accent et à l'entourage phonique. Ainsi, nous nous focaliserons sur le degré de dépendance du système de la langue seconde par rapport à la langue maternelle. Pour ce faire nous examinerons en détail la production des occlusives dans trois groupes différents : des coréens, des français et des coréens parlant français.

Nous avons constitué un corpus, dont les items sont du type V1CV2 (voyelle-consonne-voyelle), où V1 représente la voyelle/a/, tandis que V2 représente /a/ /i/. C représente les occlusives.

3. ANALYSE EXPERIMENTALE

3.1. *Résultats chez les Coréens*

Dans une première expérience I, consacrée aux aspects acoustiques et articulatoires de certaines consonnes occlusives en différents contextes, prononcées en chaîne parlée, nous effectuons des commentaires à partir de l'observation des mesures relevées pour trois locuteurs coréens. Les occlusives sourdes coréennes sont de trois types différents ; faibles, aspirées, fortes (glottalisées) pour un même lieu d'articulation (bilabiales ; p, p^h, p', alvéodentales ; t, t^h, t', vélaires ; k, k^h, k', palatales ; c, c^h, c'). Le 1er problème du trait distinctif des occlusives, c'est-à-dire, la sonorité (sonore/sourde) ou la tension articulatoire (forte/faible), a été discuté depuis une trentaine d'années par des linguistes.

Table 1. Les occlusives coréennes chez les Coréens

		Con.				Con	
	tenue	VOT	total		tenue	VOT	total
apa	-	-	53	api	-	-	67
ata	-	-	46	ati	-	-	60
aca	-	-	62	aci	-	-	88
aka	-	-	48	aki	-	-	63
ap^ha	78	38	114	ap^hi	108	92	199
at^ha	74	49	123	at^hi	75	70	144
ac^ha	118	69	187	ac^hi	83	55	138
ak^ha	89	36	125	ak^hi	65	60	125
ap'a	126	14	140	ap'i	110	20	130
at'a	95	14	109	at'i	125	19	144
ac'a	114	34	148	ac'i	127	57	184
ak'a	111	20	132	ak'i	118	28	146

(voir Figure 1 : la délimitation)

3.1.1. Etant donné que la réalisation des consonnes est la moins influencée entre deux voyelles /a/, nous donnons les différentes valeurs des consonnes aspirées par rapport aux consonnes non-aspirées, uniquement sourdes en position intervocalique.

La durée des tenues des consonnes occlusives coréennes varie en fonction de la force articulatoire : la consonne glottalisée est la plus longue, vient ensuite la consonne aspirée, et enfin la consonne faible. La durée moyenne du VOT est de 55 ms pour /p^h, t^h, c^h, k^h/ et de 25 ms pour /p', t', c', k'/. Pour /p, t, c, k/ le VOT est négatif.

Nous constatons que les durées des voyelles précédant la consonne faible sont plus brèves que celle des consonnes aspirées et glottalisées à l'exception des consonnes faibles. Car, dans ce contexte, les consonnes faibles sont entièrement sonorisées.

3.1.2. La durée moyenne de la tenue la plus longue est celle des consonnes glottalisées : 122 ms, vient ensuite celle des consonnes aspirées : 83 ms.

Par contre, le VOT a une durée moyenne de 64 ms pour /p^h, t^h, c^h, k^h/ et de 39 ms pour /p', t', c', k'/. Pour /p, t, c, k/, il y a entièrement sonorisation (Figure 2) entre les deux voyelles précédente /a/ et subséquente /i/.

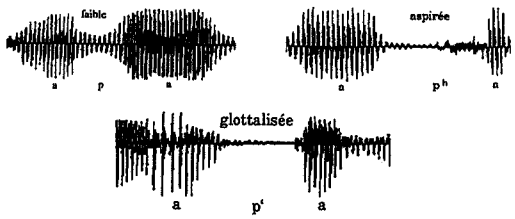


Figure 2. Signaux de /apa/, /ap^ha/ et /ap'a/ en Coréen

Dans la catégorie de la consonne aspirée, à la différence de la consonne glottalisée, nous pouvons noter qu'il y a un lien étroit entre les sourdes et les durées totales. Au fur et à mesure que la durée du VOT s'allonge, la durée totale des consonnes s'allonge.

En coréen, l'efficacité des indices ne varie pas en fonction de la réalisation phonatoire des bilabiales, des alvéodentales, des palatales et des vélaires, de sorte que nous avons regroupé les faits les concernant. Par contre, en raison du caractère particulier des trois séries (faibles, aspirées, glottalisées) qui peut modifier le degré d'efficacité de tel ou tel critère, le VOT est un indice qui permet de différencier nettement les 3 catégories d'occlusives sourdes coréennes.

3.2. *Résultats chez les Français*

Comparaison /p t k/ précédant /a/ et /i/ en français: En considérant les résultats que nous avons obtenus par la méthode instrumentale que nous avons retenue et qui nous semble particulièrement appropriée dans les explosions consonantiques, nous constatons l'existence d'un certain nombre de divergences de nature à fausser les moyennes que nous avons calculées : nous relevons en particulier une grande différence dans les mesures du VOT des consonnes occlusives devant [i] par rapport aux mesures que nous relevons devant [a].

L'incidence de la nature de la voyelle suivante sur le VOT de la consonne occlusive est assez considérable si l'on compare les tableaux récapitulatifs de mesure des occlusives devant [a] et devant [i].

Notons qu'il n'y a pas de différences dans les tendances entre les deux tableaux. C'est de façon identique que les

durées (durée totale, tenue et VOT) de /p, t, k/ précédant /i/ sont plus longues que celles de /p, t, k/ précédant /a/. Alors, la moyenne n'est pas significative, parce qu'en français il y a la palatalisation (Figure 3 ; Figure 5).

Figure 3. le /k/ français par les français

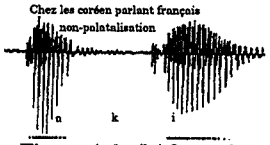


Figure 4. le /k/ français par les coréens parlant français

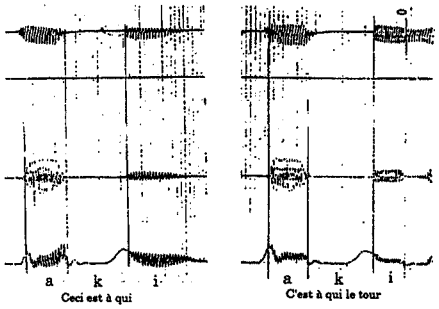


Figure 5. le /k/ palatalisé en français

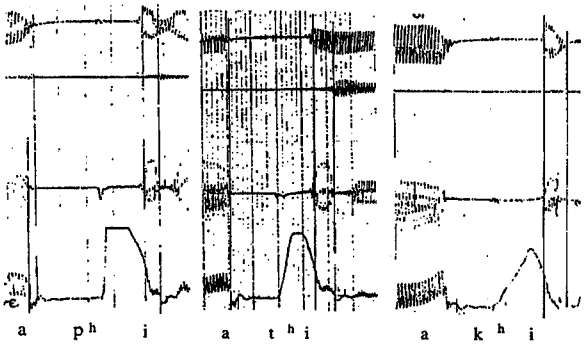


Figure 6. les occlusives aspirées du coréen

Une analyse de l'état actuel des palatalisations dans le français de nos sujets permet de constater que la vélaire se palatalise le plus facilement, que les alvéodentales et les labiales ne viennent qu'après les vélaires chez nos sujets. STRAKA explique que les vélaires /k/ se présentent comme les consonnes les moins fortes des trois séries d'occlusives, or l'énergie de la contraction des muscles élévateurs sous-linguaux, qui élèvent le corps de la langue tout droit vers le sommet de la voûte, semble pouvoir déplacer le plus facilement l'occlusion vélaire, parce qu'elle est la moins ferme de toutes (Str65).

On ne voit, dans la palatalisation consonantique, qu'une simple assimilation du lieu d'articulation d'une consonne vélaire ou alvéodentale ou alvéolaire à celui de la voyelle palatale ou du yod subséquent; cette assimilation aurait été favorisée par un accent moins énergique et par la faiblesse de la consonne (Str65).

De ce fait, et du fait également que **le système consonantique du français ne comporte pas, à la différence du coréen, des consonnes occlusives palatales, ce qui favorise une importante latitude articulatoire, puisque /k, g/ en français peuvent couvrir les zones vélaire et palatale en fonction de l'articulation qui suit.**

Table 2. Les occlusives françaises chez les Français

		Con				Con	
acc.	tenue	VOT	total	inacc	tenue	VOT	total
apa	92	17	109	apa	86	13	99
ata	86	27	114	ata	82	19	101
aka	78	31	108	aka	84	22	106
api	98	45	142	api	65	39	104
ati	66	76	141	ati	61	47	108
aki	87	70	160	aki	74	64	138

3.3. Résultats chez CF

Les consonnes /p, t, k/ sont en position accentuée par rapport aux consonnes /p, t, k/ en position inaccentuée par les Coréens parlant français. Nous suivrons la progression des occlusives sourdes entre deux voyelles /a/, et respectivement, les occlusives sourdes entre la voyelle précédente /a/ et subséquente /i/.

Table 3. Les occlusives françaises parlé par les Coréens

		Con				Con	
acc.	tenue	VOT	total	inacc	tenue	VOT	total
apa	169	20	189	apa	110	19	129
ata	149	16	165	ata	123	14	137
aka	141	30	171	aka	84	20	103
api	194	16	210	api	130	16	145
ati	167	23	190	ati	64	29	93
aki	148	33	181	aki	135	29	164

Le tableau 3 résume les données des occlusives sourdes /p, t, k/ du français en position intervocalique entre /a/ : En comparant les différentes occlusives sourdes en position accentuée et inaccentuée en entourage vocalique /a/, nous constatons que le VOT moyen des occlusives sourdes en position accentuée (21,9 ms) est également d'une longueur supérieure à celui des occlusives en position inaccentuée (17,3 ms). Quant au VOT, la longueur supérieure pour /k/ est de 30 ms, suivi de /p/ et /t/.

Les occlusives sourdes entre la voyelle précédente /a/ et subséquente /i/: Nous avons groupé dans le tableau 3 les valeurs relevées. La durée totale de /p/ est la plus longue et celle de /k/ est la plus brève entre les trois occlusives. En revanche, la durée du VOT est la plus longue pour /k/ (33 ms), suivi de /t/ (22,5 ms), et enfin de /p/ (16 ms). En ce qui concerne la réalisation de la consonne en position inaccentuée, la durée totale est la plus longue pour /k/, suivi de /p/ et de /t/. Le VOT de /t/ et /k/ est presque de même durée, et est beaucoup plus grand que celui de /p/.

4. DISCUSSION ET CONCLUSION

Il nous semble que les résultats relevés sur /p, t, k/ français nous permettent de tirer les conclusions suivantes :

Dans le même entourage vocalique, la durée de la tenue diminue à mesure que le lieu d'articulation de l'occlusive se déplace de la position antérieure vers la position postérieure. La durée de la tenue de l'occlusive bilabiale [p] est supérieure à celle de la dentale [t] qui est à son tour supérieure à celle de la vélaire [k].

Dans le même entourage vocalique, le VOT pour les occlusives sourdes s'agrandit à mesure que le lieu d'articulation de l'occlusive se déplace des lèvres vers le vélum. Le VOT de l'occlusive vélaire [k] est donc supérieur à celui de la dentale [t], tandis que le VOT est le plus bref pour la bilabiale [p]. La seule exception se présente dans le contexte vocalique de [i], où le VOT de [t] est plus long que celui de [k].

La durée de la tenue et celle du VOT dépendent du contexte vocalique. Le VOT est le plus grand quand les occlusives précèdent la voyelle [i] (Kim99). Le VOT s'avère le plus bref quand les occlusives précèdent la voyelle ouverte [a]. Le VOT est donc le plus grand pour [pi], [ti] et [ki]. Ceci est à mettre en rapport avec le caractère particulier que prend l'explosion lorsque l'occlusive est palatalisée (voir Figure 5).

Nous pouvons affirmer que les adultes sont capables de créer des phones nouveaux dans une langue seconde (Fle96 ; Kim95, 96, 99), et de modifier leurs modèles articulatoires préalablement établis lorsqu'ils produisent des phones qui leur paraissent identiques à ceux de la langue seconde. Par conséquent, il apparaît que sur le plan acoustique, les sujets coréens parlant français ont tendance à mettre dans la même catégorie des phones différents dans leur langue maternelle et dans leur langue seconde. Finalement, cela pourrait leur permettre de réaliser des phones non pas identiques mais véritablement nouveaux. Pourrait-on, via des études plus approfondies dans l'interférence linguistique chez les bilingues et dans le champ de la didactique, en arriver à explorer le concept de catégorie phonique nouvelle? Ceci nous permettrait ainsi de l'utiliser à bon escient et plus efficacement. Placées dans un contexte palatalisant (+i, j), les occlusives du coréen et du français ont été analysées selon les paramètres : VOT (voice onset time), durée de la consonne.

La voyelle [a], de plus, marque le moins les consonnes qui l'entourent, à la différence de [i] qui a tendance à provoquer la palatalisation de la consonne, surtout en français (Figure 3 ; Figure 5). Nous savons que **le français ne comporte pas de palatales sourdes, mais connaît la palatalisation.** Par contre, **le coréen comporte des palatales, mais ne connaît pas le phénomène de la palatalisation** (Figure 3 ; Figure 5).

La forme et la hauteur d'explosion varient (figure 5 ; Figure 6). En effet, la présence d'une articulation palatale subséquente donne naissance en français, au moment de l'explosion, à des bruits de friction importants dus au rapprochement du dos de la langue. L'espace entre la voûte palatine et le dos de la langue étant réduit, l'explosion s'en trouve allongée et de faible hauteur. Par contre, il y a un assourdissement des occlusives par les Coréens parlant français (Figure 2). Les occlusives sonores françaises en position intervocalique sont caractérisées par des vibrations régulières d'une amplitude stable chez les Français. Mais chez les Coréens parlant français, cette partie est le plus souvent assourdie. Même quand elle est accompagnée de vibrations laryngiennes, l'amplitude semble moins importante. De plus, cette vibration n'est pas toujours stable, et l'on constate généralement une forme décroissante. Pour les sonores françaises dites par les locuteurs coréens, il arrive même parfois que la vibration s'arrête avant l'explosion, ce qui n'arrive guère chez les sujets français. Les résultats montrent que l'amplitude, le degré de palatalisation dans les deux langues manifestent des différences systématiques. Quelle explication peut-on donner de ce comportement différent d'éléments se trouvant pourtant dans un contexte identique? La réponse à donner oblige à ouvrir des perspectives sur la notion et la portée du contexte en phonétique. Une notion théorique supplémentaire doit être prise en compte, celle du système, facteur décisif, qu'on ne peut dissocier de celle de "contrainte phonologique."

BIBLIOPRAPHIE

[Zim58] Zimmermann S.A. & Sapon S.M. (1958), "None on vowel duration seen cross linguiscally", Journal Acoust. Soc. Am., 30, 2, pp. 152-153.

[Str65] Straka G. (1965), "Naissance et disparition des consonnes palatales dans l'évolution du latin au français", Travaux de linguistique et de littérature, Ed. Klincksieck, pp. 117-168.

[Kea84] Keating P.A. (1984), "Phonetic and phonological representation of stop consonant voicing", Language 60, 286-319.

[Fle91] Flege J. (1991), "Age of learning affects the authenticity of voice-onset time (V.O.T.) in stop consonants produced in a second language", Journal Acoust. Soc. Am., 89, 1, pp. 395-411.

[Lae92] Laeufer C. (1992), "Patterns of voicing-conditioned vowel duration in French and English", Journal of Phonetics, 20, pp. 411-440.

[Kim94] Kim H.Z. (1994), "Contribution à une étude comparative des occlusives du coréen et du français", Traveaux Ins. Pho. Strasbourg. 24, 39-89.

[Kim95] Kim H.Z. (1995), "French and Korean Plosives: a Comparative Analysis", International Congress of Phonetic Sciences 95, 4. 176-179.

[Fle96] Flege, J.E., Schmidt A.M. & Wharton G. (1996) "Age of learning affects rate-dependent processing of stops in a second language", Phonetica, 53. 143-161.

[Kim96] Kim H.Z. (1996), "Quelques aspects acoustiques de la production des occlusives du coréen et du français", Journée d'étude sur la Parole 96. 21, 159-162.

[Kim99] Kim H.Z. (1999), "The production of "new" and "similar" phones in a second language", International Congress of Phonetic Sciences 99, 2, 1137-1140.

Etude phonétique (segmentale et prosodique) d'un cas de jargon phonémique

Marianne Louis * °, *Albert Di Cristo* *, *Michel Habib* * °, *Daniel Hirst* *

*Laboratoire Parole et Langage, Université de Provence, Aix-en-Provence

° Laboratoire Parole et Langage, Equipe Parole et Dyslexie, Hôpital Nord, Marseille

ABSTRACT

This study documents the segmental and prosodic competence of patient suffering from an extremely rare variety of Wernicke's aphasia with phonemic jargon. Despite the fact that the patient's speech was totally incomprehensible, in a reading task, we observed : (1) a regular correspondence between the number of phonemes produced compared to that of target text (2) a frequency distribution of phonemes equivalent to that observed for standard French (3) a conservation of prosodic characteristics reflecting both expressive and structural prosody. The results showed in general that the patient was capable of producing and reproducing the characteristic patterns of French together with their contextual variability.

1. INTRODUCTION

La présente étude s'intéresse au cas d'un patient (J.C.) présentant une aphasie de Wernicke avec jargon phonémique [Per81], qui parvient néanmoins à communiquer grâce à la prosodie et à la gestuelle.

Le patient que nous étudions révèle deux particularités intéressantes : un jargon constitué uniquement de néologismes et un site lésionnel inattendu par rapport aux résultats de l'étude clinique, ce qui constitue un cas unique en ce qui concerne le français. En effet, seuls trois cas similaires, l'un en italien [Cap94] et les deux autres en anglais américain [Per81] et [Han96], ont été décrits à ce jour dans la littérature. En outre, nous avons été frappés par la richesse et la grande variété de sa prosodie, ce qui permet d'interpréter un certain nombre d'informations relatives aux aspects les plus élémentaires de la communication. L'objectif de cette étude est d'établir un profil phonémique et un profil prosodique qui seraient révélateurs de la compétence linguistique que possède encore ce patient. Nous abordons le concept "compétence" sur la base de comparaisons établies entre les lectures d'un texte effectuées par le patient et les locuteurs d'un groupe témoin. Cette comparaison fait l'objet d'une double analyse qualitative et quantitative, à la fois au niveau segmental et au niveau suprasegmental.

2. PRÉSENTATION DU PATIENT

Le patient est âgé de 70 ans lorsqu'il souffre d'un accident vasculaire cérébral en 1994. Il se plaint de mal à la tête, tombe inconscient et, au réveil, quelques minutes plus tard, ne peut plus parler normalement. Depuis lors, il présente de façon inchangée un jargon massif et une anosognosie totale (le patient n'a pas conscience d'avoir perdu l'usage de son lexique) de ses productions. Sa compréhension orale et écrite est profondément altérée et les quelques productions écrites qu'il accepte de réaliser sont totalement jargonnées.

L'examen tomodensitométrique (figure 1) met en évidence "*un infarctus ischémique du territoire sylvien gauche qui englobe la partie antérieure du lobe temporal, la région frontale inférieure dont l'aire de Broca, l'Insula, qui s'étend en profondeur jusqu'aux ventricules mais épargne les noyaux gris et la capsule interne, et qui s'étend en haut et en arrière jusqu'à la partie inférieure du lobe pariétal.*"

Ce patient appartient au groupe des 9,7% de cas exceptionnels pour qui la localisation anatomique n'est pas corrélée au type d'aphasie [Bas85].

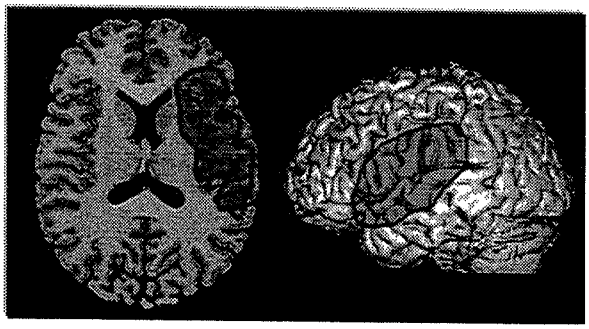


Figure1 : Représentation en gris foncé des parties lésées.

3. METHODOLOGIE

3.1 Constitution du corpus

Le patient et les locuteurs témoins ont procédé à la lecture du texte : "*Hier soir*" extrait des "*Supports verbaux en orthophonie*" de [Cel91]. Il s'agit d'une courte histoire, de six phrases, présentée en une seule partie sans paragraphes. Six lectures sont proposées au patient, les trois premières avec la consigne de lire simplement le texte et les trois dernières en précisant que chaque lecture sera accompagnée par le soulignement du doigt afin de mieux repérer les pauses et les retours à la ligne. Nous avons eu beaucoup de difficultés à identifier les productions jargonnées du patient qui ne sont pas en correspondance parfaite avec le support imposé. Nous avons seulement pu repérer des segments équivalents comme le montre l'exemple suivant :

«Hier soir, avant de s'endormir, François fumait une dernière cigarette, en relisant le cours d'allemand qu'il avait préparé pour ses élèves de terminale.»

«l ə s i t e a v ɑ̃ d a v ə d a v e d s e a f ɔ k ə f e g a ʀ a d e k ɛ̃ ʀ ɔ v a z i j e k ʀ a s i ʀ e d e v a l ɔ̃ a ʀ y s i a s e ʀ a m a a n ɔ̃ v u t i ʀ a t i ʀ a m a a s e ʀ ɔ t i l a ʀ ɔ s ɛ ʀ p y z ɛ ʀ s»

Un groupe témoin de cinq personnes, de la même tranche d'âge et de milieu socioculturel équivalent, effectue les mêmes tâches. Ces épreuves sont enregistrées et filmées avec le matériel suivant : magnétophone DAT Tascam, DA-P1 ; micro cravate Sennheiser ; laryngophone ; cassettes BASF ; caméra vidéo Blaupunkt CR-5000 ; Magnétoscope VHS-C et cassette Konoka EC-45.

3.2 Protocole expérimental

Les enregistrements ont été numérisés au moyen du logiciel PHONEDIT à la fréquence de 16Khz.

Nous avons procédé à une transcription large des six lectures jargonnées à l'aide de l'alphabet phonétique international en utilisant la police SILDoulosIPA. Pour limiter les erreurs de transcription du jargon, nous avons noté par : [ɔ] les phonèmes : [/ɔ/, /o/] ; [e] les phonèmes : [/e/, /ɛ/] ; [ə] les phonèmes : [/ə/, /œ/, /ø/] et [ɛ̃] les phonèmes : [/ɛ̃/, /œ̃/].

Cette transcription a été affinée en ayant recours aux possibilités de visualisation et d'écoute du signal offertes par le logiciel PHONEDIT que nous avons également utilisé pour procéder à l'étiquetage et à l'analyse acoustique des enregistrements.

Dans un premier temps, le corpus a été segmenté en ne considérant que les pauses communes à tous les locuteurs. Un segment de parole compris entre deux pauses est qualifié de "*segment inter-pauses*" (segment I-P). Le texte est ainsi subdivisé en 13 segments I-P.

4. RÉSULTATS

4.1 Résultats de l'étude segmentale

L'étude segmentale comporte deux étapes. La première procède d'un comptage des phonèmes, afin d'effectuer une comparaison de l'effectif des phonèmes correspondant à la transcription du texte de lecture et de celui des phonèmes de la production jargonnée du patient. La seconde étape concerne une comparaison de la fréquence d'occurrence des phonèmes de la production jargonnée avec des données de référence sur le français [Wio85]. On notera que les semi- consonnes [/j/, /ɥ/, /w/] sont comptabilisées comme des consonnes.

Le nombre de phonèmes correspondant au texte de lecture et à celui de la production jargonnée de J.C. sont respectivement égaux à : 428 et 429. De plus, pour chaque segment I-P le nombre de phonèmes contenu dans le jargon est équivalent au nombre de phonèmes déterminé par le texte.

Une comparaison des six productions jargonnées du patient montre que la distribution de phonèmes est équivalente pour chacune des lectures. La répartition globale des phonèmes du texte et celle de la production jargonnée sont superposables. On relève cependant une variabilité importante en ce qui concerne l'emploi de certains phonèmes. C'est ainsi que le jargon contient un nombre plus élevé de [/ɔ/, /a/, /u/, /ɔ̃/, /m/, /t/, /s/, /z/, /v/, /ʀ/, /w/] et un nombre plus faible de [/ə/, /ɑ̃/, /i/, /d/, /n/, /p/, /f/, /l/, /ɥ/], le nombre des autres phonèmes restant équivalent.

L'analyse individuelle des treize segments inter-pauses I.P. fait apparaître que l'inventaire des phonèmes de la production jargonnée est plus variable que celui qui est imposé par le texte. A titre d'exemple, le segment IP-13 comporte 30 phonèmes, soit 17 phonèmes différents. Le patient utilise 24 phonèmes différents, soit un tiers de plus pour construire son jargon.

En ce qui concerne la fréquence d'occurrence des phonèmes, une comparaison avec les données de référence de [Wio85] sur le français montre que la répartition des phonèmes du français, du texte et du jargon sont globalement équivalentes (figure 2).

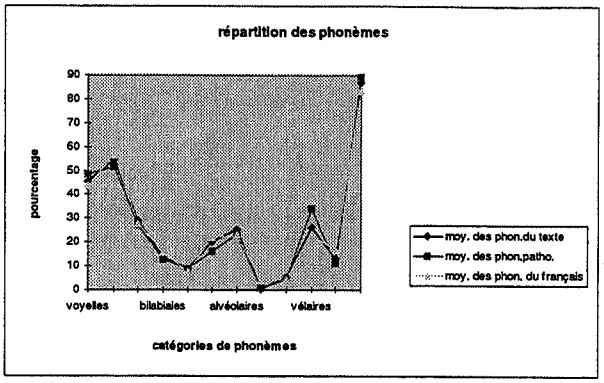


Figure2 : Distribution des phonèmes dans les 3 conditions ; jargon, texte et français selon Wioland.

Malgré cette similitude de courbes, la différence entre les phonèmes du français et les phonèmes du texte est hautement significative : p = <.0001 ; la différence entre les phonèmes du français et les phonèmes du patient est également hautement significative: p = <.0001 et la différence entre les phonèmes du texte et les phonèmes du patient est aussi significative mais de façon moindre puisque p = .0921.

Cependant, l'étude des segments I-P fait ressortir trois axes de variabilité. Premièrement, la répartition des phonèmes peut être équivalente dans les trois conditions: référence, texte et jargon; deuxièmement, elle peut être superposable seulement pour le texte et le jargon, et distincte de la référence. Enfin, elle peut être superposable pour la référence et le jargon, et distincte du texte.

4.2 Résultats de l'étude suprasegmentale

L'étude suprasegmentale concerne deux aspects de l'organisation prosodique : la distribution des proéminences associées aux accents et l'analyse des profils mélodiques des segments IP. Nous n'aborderons pas dans cette communication le premier point qui nécessiterait de longs développements.

Les profils mélodiques sont définis en termes de configurations lisses et continues correspondant à une séquence de points-cibles dont la détection est effectuée au moyen de l'algorithme MOMEL [Hir93]. Ces points-cibles font l'objet d'un codage automatique (figure 3) fondé sur l'usage de l'alphabet intonétique INTSINT [Hir98].

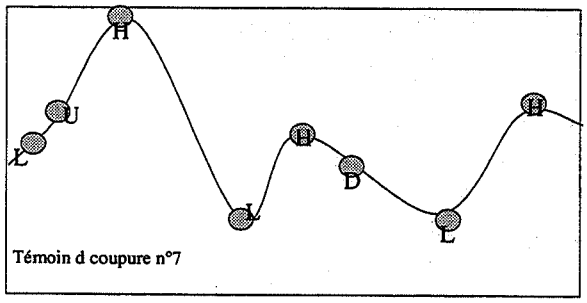


Figure3. : Illustration d'un profil mélodique et du codage INTSINT.

Dans un premier temps et afin de faciliter la comparaison des profils mélodiques, nous ne retiendrons que les points de la courbe codés L et H (soit les points bas et haut relatifs).

Nous dénombrons 16 profils mélodiques différents dans les productions des locuteurs du groupe témoin. Cet inventaire comprend des profils de faible empan, tels les profils L-H ou H-L et des profils de plus grande portée pouvant comporter des séquences de 10 cibles. Quinze profils de la production jargonnée se révèlent être similaires à ceux du groupe témoin.

Un examen longitudinal des profils associés aux segments inter-pauses montre que la variabilité séquentielle des profils est plus importante dans la production jargonnée au cours de la lecture de la première partie du texte, et que cette variabilité tend à se stabiliser vers la fin du texte, ce qui donne l'impression de la mise en place d'une génération mécanique des profils.

Il est particulièrement intéressant d'observer que, du point de vue qualitatif, la production du patient met en évidence un nombre plus élevé de profils, ce qui correspond à une plus forte récursivité de segmentation intonative du texte. Mais, fait plus remarquable encore, la fréquence d'occurrence du profil simple L-H est significativement plus élevée chez le patient que chez les locuteurs du groupe témoin (40% de l'ensemble des profils du texte, contre 11%). Or, le profil L-H est le patron de base du système intonatif du français, comme l'admet la majorité des modèles qui s'intéressent à la phonologie prosodique de cette langue. Il est donc légitime de fonder l'hypothèse que la compétence prosodique profonde du patient n'est pas annihilée et que son handicap vient surtout de son incapacité à gérer fonctionnellement de façon systématique ce substrat phonologique. Des recherches en cours ont pour objet de vérifier le bien fondé de cette hypothèse.

5. DISCUSSION

Le patient dont nous avons entrepris l'étude présente un type rare d'aphasie, l'aphasie de Wernicke avec jargon phonémique. Il représente de surcroît une exception dans les règles de la corrélation anatomo-clinique entre le siège anatomique de la lésion et le type d'aphasie : en effet sa production est fluente malgré une destruction de l'aire de Broca.

La surprenante préservation de la qualité mélodique de son discours pourrait entretenir chez lui l'illusion que son langage est normal expliquant l'intensité de l'anosognosie. Les investigations préliminaires des caractéristiques segmentales et suprasegmentales de sa parole jargonnée auxquelles se limitent cette étude permettent de constater que dans toutes les épreuves, à un moment ponctuel [Jak69], ce patient démontre des capacités de reproduction parfaite. Par conséquent, il apparaît chez ce patient que des patrons tant phonémiques que prosodiques sont non seulement intacts mais encore accessibles.

En revanche, à l'intérieur de chacun de ces deux systèmes, le patient semble utiliser les éléments à sa disposition comme le type et le nombre de phonèmes et le type et le nombre de profils mélodiques, de manière non systématisée, en tout cas non conforme au code lexical et au contenu sémantique. Malgré ces divergences, nous remarquons un respect des patrons de base du français, dans la mesure où les similitudes structurelles transcendent la variabilité inhérente à la parole.

Nous faisons l'hypothèse que l'organisation cérébrale atypique de ce patient intervient dans la singularité de son tableau clinique. Sa production articulatoire est probablement sous la dépendance des régions motrices de l'hémisphère droit, tant du point de vue phonétique que prosodique. L'aire de Broca gauche au contraire, paraît dénuée de tout rôle articulatoire.

6. CONCLUSION

L'étude segmentale met en évidence une conservation des unités phonémiques, une programmation du nombre de ces unités conforme au modèle imposé et une distribution aléatoire des unités. L'étude suprasegmentale révèle une conservation des unités prosodiques, une programmation du nombre des profils mélodiques conforme au modèle imposé et une distribution aléatoire de ces profils mélodiques. Les paramètres conservés et les paramètres altérés sont les mêmes au niveau segmental et au niveau suprasegmental. Parmi les distributions aléatoires, nous retrouvons sur les deux niveaux d'étude la possibilité d'obtenir ponctuellement un résultat strictement conforme au modèle imposé, ce qui dénote une prégnance des codes phonémiques et prosodiques, mais une incapacité à les gérer d'une façon systématique.

BIBLIOGRAPHIE

[Bas85] Basso A. (1985) "Anatomoclinical correlations of the aphasias as defined through computerized tomography : exceptions.", Brain and language, Vol. 26, pp. 201-229.

[Cap94] Cappa S. (1994) "Case study: glossolalic jargon after a right hemispheric stroke in a patient with Wernicke's aphasia", Aphasiology, Vol. 8, pp. 83-87.

[Cel91] Celerier P. (1991) Supports verbaux en orthophonie, Isbergues.

[Han96] Hanlon R. (1996) "Disconnected phonology: A linguistic analysis of phonemic jargon aphasia", Brain and Language, Vol. 55, pp. 199-212.

[Hir93] Hirst D., Espesser R. (1993) "Automatic modelling of fundamental frequency curves using a quadratic spline function", T.I.P.A., Vol. 15, pp. 71-85.

[Hir98] Hirst D., Di Cristo A. (1998), Intonation Systems. Cambridge University Press.

[Jak69] Jakobson R. (1969), Langage enfantin et aphasie, Paris : Editions de minuit.

[Per81] Perecman E. (1981), "Phonemic jargon : a case report ", in Jargonaphasia, Ed J Brown (New York, London, Toronto, Sydney, San Francisco) pp. 177-259.

[Wio85] Wioland F. (1985), Les structures syllabiques du français, Genève-Paris : Slatkine-Champion.

Peigne et brosse pour Fo :
Mesure de la fréquence fondamentale par alignement de spectres séquentiels

Philippe Martin

Département d'Études Françaises, Université de Toronto
Carr Hall, St Joseph St.
Toronto, Ontario, Canada M5S 1J4
Tél.: ++1 416 960 6122 - Fax: ++1 416 920 4634
Mél: pmartin@chass.utoronto.ca
http://www.chass.utoronto.ca/french/ling/Homepage/martin.html

ABSTRACT

The spectral comb method to measure the fundamental frequency of speech signal was introduced in 1981. Although very robust in principle, as it ensures the correct detection of F0 even if some harmonics are missing in the signal, its implementation requires care in order to obtain satisfactory results. In this paper we discuss various aspects of the process, and we introduce a novel method for better selection of harmonics in the short-time spectrum before the spectral comb is computed. This method, called the spectral brush, aims for a better reduction of the effect of noise as well as the detection of Fo of simultaneous speech sources.

1. LA MÉTHODE DU PEIGNE SPECTRAL

1.1 Principe

La méthode d'estimation de la fréquence laryngienne par le peigne spectral procède par intercorrélation du spectre à court terme du signal avec une fonction « peigne » constituée d'une série d'impulsions de Dirac espacées de la fréquence fondamentale recherchée [Mar81]. La présence d'une structure harmonique dans le signal se traduit par un maximum de la fonction d'intercorrélation lorsque l'intervalle entre les impulsions successives correspond à la fréquence fondamentale.

Si A_n représente les amplitudes des composantes du spectre à court terme et δ_n les amplitudes des harmoniques du peigne espacées de la fréquence f, la fonction peigne P(f) s'écrit

$$P(f) = \sum_{1}^{F_{max}/f} \delta_n A_n$$

avec F_{max}/f représentant le nombre d'harmoniques prises en compte dans le calcul, et liées à la fréquence du spectre maximale utilisable F_{max}. Cette méthode peut être considérée comme la recherche d'un maximum de probabilité de la valeur de Fo à partir d'un spectre donné [Kon94].

1.2 Sous-harmoniques

Pour éviter la détection de sous-harmoniques, les amplitudes des dents du peigne reçoivent une amplitude décroissante avec la fréquence. Une fonction décroissante monotone sur les dents du peigne garantit que l'amplitude des sous harmoniques sera inférieure à celle de la fondamentale :

$$P(f_0/k) < P(f_0) \text{ quel que soit k}$$

pour autant que le spectre soit effectivement constitué d'impulsions de Dirac sur l'axe des fréquences. Les valeurs intermédiaires des sous-harmoniques du peigne correspondent alors à des valeurs nulles.

En réalité, le spectre à court terme d'un segment voisé du signal de parole ne présente pas cette propriété et les intervalles entre les pics harmoniques ne sont pas nuls (fig. 1).

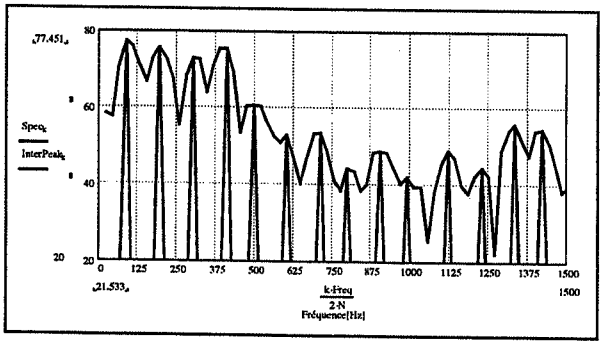


Figure 1 : spectre de la voyelle [a], fenêtre de Hanning de 512 points à 11025 Hz

Pour se rapprocher d'un spectre à plus grande résolution fréquentielle, on procède alors à la discrétisation du spectre à court terme.

1.3 Discrétisation du spectre

Le signal de parole étant essentiellement non-stationnaire, l'augmentation de la durée de la fenêtre pour obtenir une meilleure résolution fréquentielle ne peut guère dépasser 30 à 50 ms. Une meilleure résolution fréquentielle des harmoniques ne peut donc être obtenue qu'au prix d'une perte de résolution temporelle inacceptable. Procédant par analogie avec un modèle de l'audition intégrant l'effet de masque [Kor68], on procède à une discrétisation du spectre à court terme à partir des sommets détectés dans A_n pour lesquels $A_{k-1} < A_k$ et $A_k >= A_{k+1}$ Une interpolation parabolique permet d'estimer la fréquence et l'amplitude de chaque harmonique à partir de A_{k-1}, A_k et A_{k+1} (cette interpolation n'est correcte du reste que si la fenêtre de prélèvement du signal est une gaussienne). On obtient ainsi un ensemble de valeurs permettant de reconstituer un spectre discrétisé à haute résolution, constitué pour chaque harmonique d'une parabole de largeur de bande ΔW Hz à –20 dB sur une échelle de fréquence à ΔF Hz de résolution (valeurs typiques : ΔW = 30 Hz et ΔF = 1 Hz).

1.4 Les dents du peigne

La prise en compte des harmoniques d'ordre élevé pose un autre problème. Le calcul de l'intercorrélation entre le spectre discrétisé et une fonction peigne dont l'espacement entre les dents varie par sauts de ΔF Hz, peut donner des résultats erronés du fait de l'erreur entraînée par l'utilisation d'entiers dans le calcul. En effet, un saut de 1 Hz correspond à une variation de n Hz pour la n^{ème} harmonique. Dans ce cas, l'intercorrélation pour n = 50 par exemple peut toucher aussi bien la 50ème que la 51ème harmonique d'un signal de fondamentale de 100 Hz.

L'ordre supérieur $H_{\max}$ de l'harmonique utile dans le calcul du peigne est donc limité d'une part par la largeur de bande ΔW des pics harmoniques, et de l'autre par la résolution fréquentielle ΔFo adoptée pour le calcul de la fonction peigne, avec $H_{\max} = \Delta W / \Delta F$. Les valeurs retenues plus haut donnent une limite de 30.

1.5 Le peigne rapide

Le calcul direct de la fonction peigne mène à un grand nombre de termes nuls, du fait des valeurs nulles entre les pics du spectre discrétisé. Le nombre de multiplications additions nécessaires pour H harmoniques est donné par

$$\sum_{F\min}^{F\max} \sum_{1}^{H} \delta_n A_n \text{ avec } H = F_{Max} / f$$

Un résultant équivalent est obtenu en additionnant n spectres discrétisés, transformés linéairement en fréquence par le facteur 1/n et en amplitude par la fonction de décroissance des dents du peigne δ_n [Mar86]. Le nombre d'opérations (additions-multiplications) est cette fois réduit à $n(F_{Max} - F_{Min}) / \Delta F$.

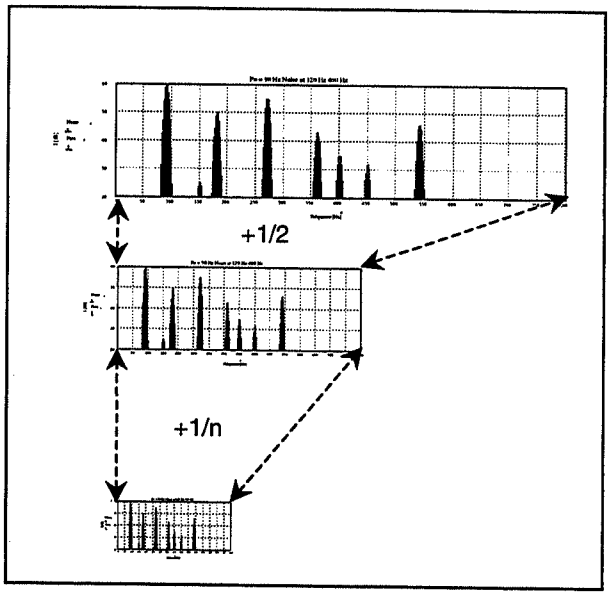


Figure 2 : Calcul du peigne rapide par addition de n spectres réduits en fréquence par n et en amplitude par les dents δ_n du peigne.

En effectuant les calculs à partir d'une table contenant les fréquences et les amplitudes des harmoniques détectées, le calcul du peigne se limite à additionner des paraboles de ΔW points réduites en amplitude par le coefficient des dents du peigne δ_n. Le nombre total d'opérations est alors réduit à $nH(\Delta W / \Delta F)$.

1.6 Le peigne différentiel

Pour utiliser les informations apportées par les harmoniques d'ordre élevé, on peut également intégrer dans le calcul du peigne les différences entre sommets du spectre, différences d'ordre $1,2,...,m$. Chaque différence entre sommets d'ordre d est représentée par une parabole comme dans le cas du peigne direct, qui apparaît ainsi comme un peigne d'ordre 0. L'amplitude de cette parabole est égale à la moyenne des amplitudes des sommets impliqués dans la différence. Ctte implémentation est réalisée dans WinPitch [Win96].

2. LA BROSSE SPECTRALE

2.1 De meilleures harmoniques

Plusieurs procédés heuristiques sont couramment employés pour améliorer la qualité du spectre à court terme, pour obtenir des harmoniques représentatives des sons voisés. On peut par exemple aplatir le spectre en renforçant les amplitudes des harmoniques selon la courbe de réponse inversée du filtre correspondant au conduit vocal, On peut aussi ne retenir que les harmoniques présentant une différence supérieure à un certain seuil par rapport aux vallées avoisinantes, etc.

Ces procédés donnent des résultats plus ou moins satisfaisants selon la nature du signal analysé, et surtout selon la nature du bruit, le bruit étant défini (implicitement) comme tout effet spectral de ce qui n'est pas rendu compte dans un modèle source-filtre de la phonation.

2.2 Séparation des harmoniques

Un cas classique de bruit est constitué par la présence d'un ou de plusieurs autres signaux de parole dans le signal analysé (« cocktail party ») ou de tout autre signal à structure harmonique (accompagnement musical par exemple). De nombreuses approches ont été proposées pour résoudre ce problème [cf. Dov93]. Nous proposons ici une méthode, exploitant (comme dans le cas du peigne) les propriétés de structure harmonique des sons voisés de la parole, propriétés conservées à l'intérieur d'un segment voisé.

Contrairement au spectre du signal de parole dans ses parties voisées à structure harmonique, les « bruits » à structure harmonique aléatoires, stationnaires ou musicaux présentent des caractéristiques spécifiques qui permettent de les différentier des spectres du signal de parole. On peut distinguer:

a) Les bruits transitoires (chocs,...);
b) Les bruits stationnaires (moteur,...);
c) Les bruits musicaux (à hauteur constante).

Pour cette dernière catégorie, la grande majorité des instruments de musique présentent des séquences de spectres de fréquence fondamentale stationnaire (les glissandos de fréquence y sont rares).

Toutes ces sources ont des caractéristiques de structure harmonique différentes des segments voisés du signal vocal. Mis à part le cas de la voix chantée (et seulement dans l'hypothèse d'une stationnarité parfaite de la fréquence laryngée dans le chant), la fréquence fondamentale est essentiellement variable au cours du temps. Le principe de la brosse spectrale est d'alors de séparer les harmoniques des sons voisés qui évoluent dans le temps proportionnellement à la fondamentale des harmoniques de celles d'un bruit transitoire, stationnaire ou musical en corrélant entre eux des spectres à court terme successifs.

2.3 Alignement des harmoniques

Soient 2 spectres à court terme discrétisés aux instants t et t+1, espacés d'une demi-durée de fenêtre de prélèvement du signal. Dans la partie voisée, la variation maximale de fréquence fondamentale (hors transition à un registre falsetto) est d'environ 1 % de Fo par ms [Sun79]. On effectue alors l'intercorrélation du premier spectre avec le second dont on varie linéairement l'échelle des fréquences dans l'intervalle de variation attendu (par exemple 0,80 – 1.20 si les spectres sont espacés de 20 ms).

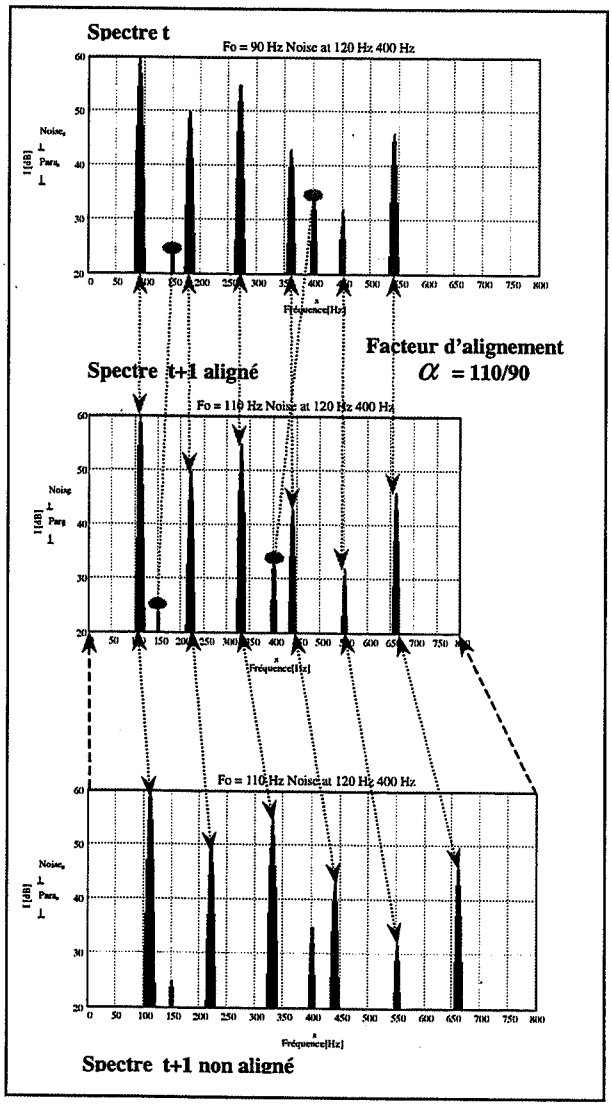


Figure 3 : La brosse spectrale. L'alignement des harmoniques par intercorrélation de spectres successifs aux instants t et t+1 permet l'élimination des pics de bruits (ici à 150 Hz et 400 Hz).

Un maximum de la fonction sera obtenu lorsque les harmoniques seront mises en correspondance pour une valeur de facteur d'alignement α correspondant au rapport des fréquences fondamentales des spectres F_t et F_{t+1} aux instants t et t+1, soit

$$\text{Max de } B(\alpha) = \sum_{f=0}^{\alpha F_{Max}} Ft(f) F_{t+1}(\alpha f)$$

Les pics qui ne sont pas en correspondance entre les 2 spectres sont alors éliminés. Ils correspondent soit à du bruit transitoire, soit à des harmoniques d'une fréquence fondamentale stationnaire entre les 2 spectres, mais qui de se trouvent plus alignés.

Ce processus constitue une solution sub-optimale du problème global d'alignement des spectres sur l'ensemble d'une séquence voisée du signal, dans lequel il faut trouver la séquence optimale des facteurs d'alignement de chaque spectre. En procédant par spectres contigus, on détermine une solution localement optimale.

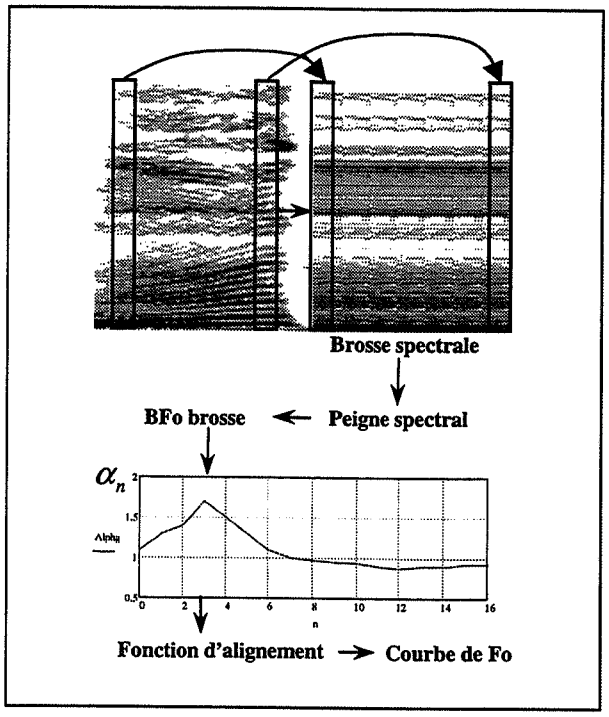


Figure 4 : La brosse spectrale aligne les harmoniques des spectres successifs sur celles du premier spectre du segment voisé. Le peigne spectral détermine une seule valeur de Fo à partir de l'ensemble des harmoniques alignées. La courbe de Fo du segment est donnée par $Fo_n = BFo / \alpha_n$, avec α_n valeur de l'alignement pour les spectres successifs aux instants t. La courbe de Fo est reconstituée à partir de Fo global et de la courbe d'alignement.

La méthode de la brosse spectrale procède par alignement des harmoniques de spectres successifs, d'où son nom de « brosse spectrale ». Elle trouve tout naturellement une extension à la mesure de Fo d'un signal de parole multi-source, lorsque plus d'un maximum de la fonction brosse B(a) est retenu. Un algorithme de programmation dynamique peut alors être utilisé pour déterminer les chemins optimaux pour chacune des structures harmoniques présentes dans le signal.

2.4 Implémentation

La méthode de la brosse spectrale est présentement implémentée dans la nouvelle version du logiciel d'analyse prosodique WinPitch [Mar96], [Win96]. Des tests sur un grand nombre de types de voix et de présence de bruit est actuellement en cours.

BIBLIOGRAPHIE

[Dov93] Doval Boris, Rodet, Xavier (1993) "Fundamental Frequency Estimation and Tracking using Maximum Likelihodd Harmonic Matching and HMMs", Proc. IEEE-ICASSP 93. 221-224

[Kon94] Kondoz, A.M. (1994) Digital Speech, Wiley, Chinchester.

[Kor68] Korn, T.S. (1968) "La Notion de Fréquence du Son, Acustica", Vol. 20, No1, 55-61.

[Mar81] Martin, Ph. (1981) "Mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne", *Actes des XIIèmes JEP*, Montréal, 1981.

[Mar86] Martin, Ph. (1986) "A Fast Spectral Comb Algorithm for Fo Detection", Proceedings from the 12 International Congress of Acoustics, Toronto, Paper A6-9.

[Mar96] Martin, Ph. (1996) "WinPitch : un logiciel d'analyse temps réel de la fréquence fondamentale fonctionnant sous Windows", Actes des XXIV Journées d'Etude sur la Parole, Avignon, mai 1996, 224-227.

[Sun79] Sunberg, J. (1979) "Maximum speed of speech changes in singers and untrained subjects", Journal of Phonetics, 71-79.

[Win96] WinPitch : http://www.winpitch.com/

Contours intonatifs de la phrase interrogative en arabe

A. Zaki [1], A. Rajouani [2], M. Najim [1]

[1]Equipe Signal et Image, ENSERB. B.P 99, F-33 402 Talence Cedex, France
[2]LEESA, Faculté des Sciences. B.P 1014 - Rabat, Maroc

Tél.: ++33 556 84 61 85 - Fax: ++33 556 84 84 06
Mél: zaki@tsi.u-bordeaux.fr, rajouani@fsr.ac.ma

ABSTRACT

This paper deals with the improvement of the quality of an Arabic TTS system i.e. interrogative intonation. A set of 80 verbal interrogative sentences covering the usual interrogation strategies is analyzed. The natural F0 curves are stylized in order to constitute a database of "close-copies". Then a set of rules, which are syntax independent, are established to describe the variations of the stylized F0 curves. Despite the wide variability in F0 contours, we propose to classify them in two categories. The proposed rules are incorporated in Arabic TTS system. Interrogative sentences synthesized is relatively natural.

1. INTRODUCTION

Ce travail s'inscrit dans le cadre d'une étude de la prosodie de l'arabe en vue d'améliorer le naturel d'un système de synthèse à partir du texte [Raj96]. L'étude porte sur l'analyse des variations de la fréquence fondamentale F0 dans le contexte de la formulation d'un modèle intonatif de la phrase interrogative verbale dans ses différentes réalisations.

L'ensemble des règles de calcul des courbes intonatives doit permettre un traitement entièrement automatique et indépendant des informations syntaxiques. La validation du modèle est jugée au niveau perceptif.

2. ANALYSE ET STYLISATION

Le corpus étudié est constitué de 80 phrases interrogatives couvrant un spectre très large de structures syntaxiques de l'arabe. Les enregistrements sont effectués dans une salle sourde. La lecture du corpus est faite par un locuteur marocain. Les signaux ont été numérisés à une fréquence d'échantillonnage de 10 KHz avec une résolution de 16 bits.

Le corpus contient des phrases interrogatives englobant :

- les phrases sans marqueur d'interrogation ou avec le marqueur /hal/ et /^a/. Ces types d'interrogation admettent la réponse oui/non (yes/no questions) ;
- les phrases avec les marqueurs d'interrogation: /man/(qui), /maadaa/ (qu'est ce que), /mataa/ (quand), /kayfa/ (comment), /limaadaa/ (pourquoi), /^ayna/ (où).

Compte tenu de la complexité d'interprétation des courbes intonatives brutes, due en partie aux variations d'ordre microprosodique, il est nécessaire de réaliser une stylisation selon le sens proposé dans [Col90].

Le principe de stylisation est fondé sur l'hypothèse qu'un certain nombre d'événements présents dans le contour intonatif brut peuvent ne pas être pris en compte sans changement au niveau perceptif. Le but de la stylisation de la courbe de F0 est d'obtenir un contour intonatif réduit (ou squelettique) équivalent au niveau perceptif au contour naturel [Bea94]. La méthodologie adoptée requiert l'utilisation d'un système d'analyse par synthèse. Les copies exactes stylisées sont obtenues de manière séquentielle et sont composées d'un nombre minimal de segments. Le calcul des contours intonatifs bruts est réalisé au moyen d'un logiciel, et la stylisation se fait par le biais de son interface graphique [Mar97]. La figure 1 illustre la stylisation d'une phrase interrogative avec le marqueur /^ayna/.

3. EVALUATION PERCEPTIVE DES COPIES EXACTES

Pour valider la pertinence de la stylisation, il est indispensable de réaliser une série de tests perceptifs sur des sujets « neutres ». La procédure de test adoptée a été notamment utilisée dans le cas du français [Bea94] et de l'anglais [Pij83].

Pour cette étape, 40 phrases sont choisies parmi les 80 disponibles. Ces phrases sont sélectionnées en fonction de leur longueur (2 à 7mots). Il est en effet utile de choisir des phrases relativement courtes de manière à faciliter la mémorisation de l'intégralité de l'intonation par les sujets, et obtenir ainsi des jugements plus précis.

Pour chacune des phrases, 4 catégories constituées chacune d'une paire de phrases sont considérées :

Catégorie A : deux fois l'original resynthétisé ;

Catégorie B : original resynthétisé et copie exacte ;

Catégorie C : copie exacte et 1ère version alternative ;

Catégorie D : copie exacte et 2ème version alternative.

Pour les catégories A et B, nous utilisons l'original resynthétisé[1] plutôt que la phrase naturelle, en raison de la différence de qualité induite par le synthétiseur. En effet la dégradation de la qualité due à la synthèse est plus aisée à détecter qu'une modification d'intonation.

Pour les catégories C et D, des versions alternatives sont utilisées. Elles sont obtenues à partir des « copies exactes », par des modifications touchant la pente d'un mouvement, la hauteur d'un niveau intonatif et le timing (le mouvement est déplacé le long de l'axe temporel d'un phonème pour la catégorie C et deux pour la catégorie D). La première version alternative comporte des modifications fines par rapport à la copie exacte, alors que la seconde contient des variations plus marquées.

L'ordre de présentation des stimuli est aléatoire. Chaque sujet doit répondre par oui ou non à la question : « *Les deux mélodies d'une paire sont-elles identiques?* ». Les résultats obtenus sur les 14 sujets qui ont participé à l'évaluation, sont présentés dans la table 1.

Table 1 : Résultats des tests d'évaluation perceptive.

CAT.	NRT	NRO	MOY NRO %	NRN	MOY NRN %	EC %
A	2240	1990	88.84	250	11.16	9.62
B	2240	1902	84.90	338	15.10	9.54
C	2240	1546	69.00	694	31.00	16.00
D	2240	653	29.17	1587	70.83	15.44

NRT : nombre total de réponses, NRO : nombre de réponses « Oui », MOY NRO% : pourcentage de réponses « Oui », NRN : nombres de réponses « Non », MOY NRN% : pourcentage de réponses « Non », EC% : écart type en pourcentage de réponses « Oui ».

Pour la catégorie A, le résultat devait être en principe égal à 100%, puisque les deux éléments de la paire sont identiques. Nous obtenons cependant un score de 88.84% avec un écart-type de 9.62%.

Concernant la catégorie B, le pourcentage de paires perçues comme constituées de stimuli identiques est de 84.90%, soit une différence de 3.94% par rapport à la catégorie A (l'écart-type restant du même ordre 9.54%).

[1] Il s'agit d'une technique de synthèse semblable à PSOLA, basée sur la répétition de parcelles de signal pitch synchrone [Mar97].

Ces résultats nous permettent de conclure que les copies exactes constituent une référence relativement acceptable pour dégager les faits intonatifs pertinents pour la phrase interrogative verbale en arabe.

4. DESCRIPTION ET INTERPRETATION DES COPIES EXACTES

L'analyse des contours intonatifs des copies exactes permet de distinguer deux stratégies pour la réalisation de l'interrogation :

4.1 Première stratégie

Cette première stratégie est décrite par un contour intonatif constitué de deux composantes intonatives principales. Ce contour est caractéristique des phrases sans marqueur d'interrogation ou ayant les marqueurs /^a/ ou / hal/ (yes/no questions).

- La première composante comprend au minimum deux mouvements mélodiques, le premier montant et le deuxième descendant.
- La seconde composante est formée au maximum de deux mouvements mélodiques, le premier montant et le second descendant. Cependant, dans la majorité des cas nous n'avons qu'un seul mouvement montant qui se réalise sur le dernier mot de la phrase et qui atteint son maximum sur la dernière syllabe de la phrase à un niveau intonatif supérieur ou égal au maximum de la première composante.

Table 2 :Les quatre tableaux ci-dessous représentent les variations des mouvements intonatifs de quelques phrases interrogatives en fonction de leurs structure syllabique.

Variations mélodiques	↑ ↓ ↑↓
Phrase	D A R A B A L W A L A D U ^ A X A A H U L J A M I I L A ?
Syllabification	C V C V C V C C V C V C V C V C V V C V C C V C V V C V

Variations mélodiques	↑ ↓ ↑
Phrase	^ A D A R A B A L W A L A D U ^ A X A A H U ?
Syllabification	C V C V C V C V C C V C V C V C V C V C V V C V

Variations mélodiques	↑ ↓ ↑↓
Phrase	^ A D A R A B A L W A L A D U ^ A X A A H U S S A P I I R ?
Syllabification	C V C V C V C V C C V C V C V C V C V C V V C V C C V C V V C

Variations mélodiques	↑ ↓ ↑
Phrase	H A L D A R A B A L W A L A D U ^ A X A A H U ?
Syllabification	C V C C V C V C V C C V C V C V C V C V C V V C V

Les règles retenues pour ce type d'interrogation (yes/no questions) sont :

- Le maximum intonatif de la phrase se réalise sur le marqueur d'interrogation pour le cas de /hal/. Dans le cas de /^a/, c'est la première syllabe du verbe qui suit le marqueur qui porte le maximum intonatif. Pour la phrase interrogative sans marqueur d'interrogation, les règles d'accentuation lexicales sont vérifiées [Raj89] et le maximum intonatif se réalise sur la syllabe accentuée du verbe.
- La phrase se termine par un mouvement montant qui se réalise sur le dernier mot à un niveau supérieur ou égal à celui du maximum antécédent.
- Le minimum intonatif se réalise au début de la première syllabe du dernier mot de la phrase.
- Le contour intonatif de la phrase comprend au minimum trois mouvements mélodiques, deux montants et un descendant.

4.2 Seconde stratégie

Un second type de contour intonatif qui caractérise les phrases dont le marqueur d'interrogation est soit un pronom interrogatif (/man/, /mataa/), soit un adverbe d'interrogation (/kayfa/, /^ayna/, /limaadaa/, /maadaa/). Ce contour est formé, au minimum, de deux mouvements mélodiques qui se réalisent sur l'ensemble de la phrase du début à la fin. Le premier mouvement est montant, le deuxième est descendant. Le nombre de mouvements peut dans certains cas être supérieur à deux en respectant la tendance générale de la fréquence fondamentale à décroître lentement du début jusqu'à la fin de la phrase.

Table 3 : Les cinq tableaux ci-dessous représentent des exemples des variations des mouvements mélodiques pour le type d'interrogation considéré.

Variations mélodiques	↑ ↓
Phrase	^AYNADARABALWALADU^AXAAHU?
Syllabification	CVCCVCV CVCVCCV CVCVCVCVVCV

Variations mélodiques	↑ ↓ ↑ ↓
Phrase	KAYFADARABALWALADU^AXAAHU?
Syllabification	CVCCVCVCVCVCC VCVCVCVCVVCV

Variations mélodiques	↑ ↓ ↓
Phrase	MATAABADA^ALMU>ALLIMUDDARSA?
Syllabification	CVCVVCVCVCVCCVCVCCVCVCCVCCV

Variations mélodiques	↑ ↓ ↓ ↓
Phrase	LIMAADAADARABALWALADU^AXAAHU?
Syllabification	CVCVVCVVCVCVCVCCVCVCVCVCVVCV

Variations mélodiques	↑ ↓
Phrase	MANJAA^AFISSABAAH IMA>A^A B I IKA?
Syllabification	CVCCVVCVCVCCVCVVCVCVCVCVCVVCV

Les règles retenues pour les phrases interrogatives avec les marqueurs d'interrogation /kayfa/, /^ayna/, /mataa/, /limaadaa/, /man/ et /maadaa/ sont :

- Le maximum intonatif de la phrase se réalise sur la syllabe accentuée du marqueur d'interrogation.
- Le minimum intonatif se réalise sur la dernière syllabe du dernier mot de la phrase.

5. VALIDATION DES REGLES PAR SYNTHESE

Après avoir établi l'ensemble des règles qui décrivent les variations intonatives de l'interrogation, une validation par synthèse est nécessaire pour quantifier les variations quantitatives et apprécier l'amélioration apportée au naturel de la parole synthétisée.

Pour cela, nous procédons à générer des phrases de synthèse par le TTS en tenant compte des règles retenues.

Nous travaillons sur 30 phrases représentatives. Pour chaque phrase, on propose cinq contours intonatifs avec des niveaux intonatifs et des pentes variables. Par ailleurs, nous conservons les durées phonétiques calculées automatiquement par le système.

Les phrases de synthèse ainsi obtenues ont subi un test de préférence impliquant 9 sujets. Les valeurs retenues et correspondant au meilleurs scores de préférence sont présentées dans la table 4.

Les figures 2 et 3 présentent des exemples de contours retenus, pour chacune des deux stratégies d'interrogation et qui ont été jugés appréciables.

L'exploitation des résultats obtenus permet une amélioration marquante du naturel de la parole synthétisé.

Table 4 : Résultats retenues (les valeurs sont en Hz)

	Attaque intonative	**1er Max. intonatif**	**Min. intonatif**	**2ème Max. intonatif**
1er stratégie	130	160	136	165
2ème stratégie	135	160	100	N'existe pas

6. CONCLUSION ET PERSPECTIVES

Dans ce travail nous avons présenté les éléments essentiels pour le calcul automatique des contours intonatifs de la phrase interrogative verbale arabe. Les règles proposées ne nécessitent aucune information syntaxique. Les tests de préférence ont montré l'amélioration du naturel apporté par le traitement proposé. Cependant un réglage des variations quantitatives du contour intonatif en fonction du rythme serait souhaitable.

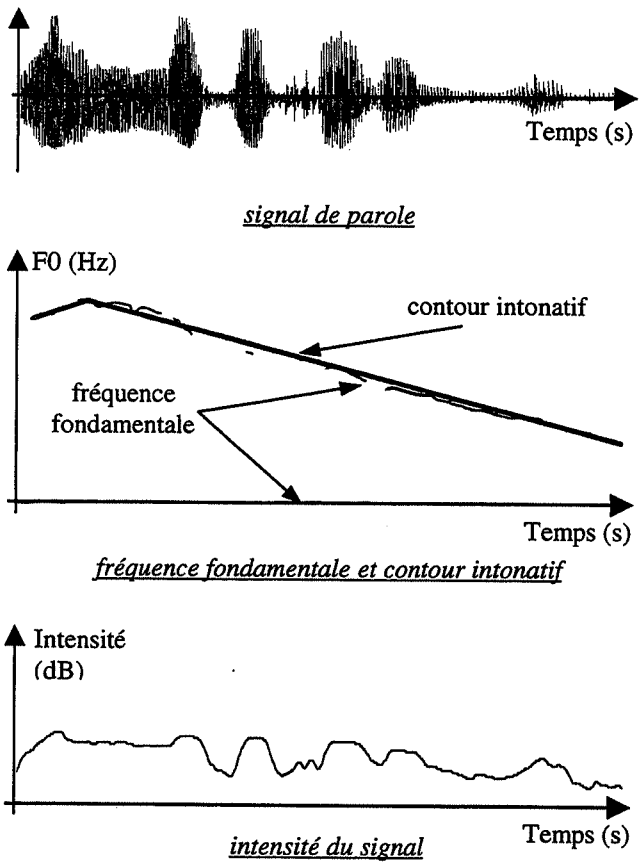


Figure 1 : Exemple de l'étape de stylisation

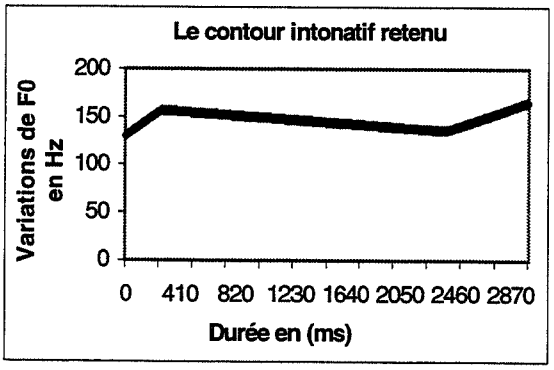


Figure 2 : schéma du contour intonatif obtenu pour les phrases interrogatives sans marqueur ou avec les marqueurs d'interrogation /^a/ et /hal/.

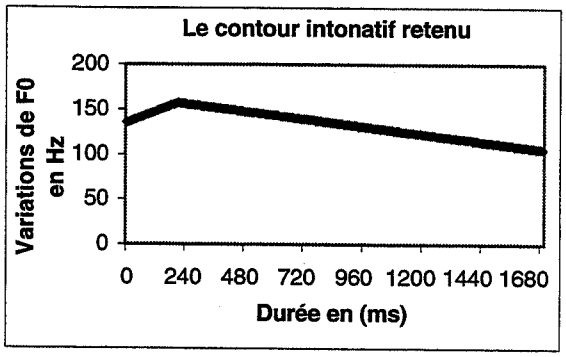


Figure 3 : schéma du contour intonatif obtenu pour les phrases interrogatives avec les autres marqueurs d'interrogation.

BIBLIOGRAPHIE

[Raj96] Rajouani A. et al. (1996), "An Arabic text-to-speech system based on rules", Proceedings. 5th ICEMCO'96, Cambridge, pp : 65-69.

[Col90] Collier R. (1990) "On the Perceptual Analysis of Intonation", Speech Communication, Vol.9, 5/6, pp :443-451.

[Bea94] Beaugende F. (1994) "Une étude perceptive de l'intonation du français : Développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte". Thèse de Doctorat, Université Paris XI-Orsay.

[Mar97] Martin P. (1997), Winpitch, logiciel d'analyse temps réel de la fréquence fondamentale, http://www.winpitch.com.

[Pij83] de Pijper J.R. (1983) "Modelling British English intonation", Foris Publications, Dordrecht.

[Raj89] Rajouani A. et al (1989) "Etude de l'accent lexical en arabe", Journal d'Acoustique, 2, pp : 171-176.

Expressions prosodiques de certaines attitudes en tchèque et en français

Jana Mejvaldová

Institut de Phonétique, Université Charles,
nám. J. Palacha 2, 116 38 Prague 1, République Tchèque
tél.: ++4202 216 19 250 – fax: ++420 24 81 21 66
mejvaldo@ff.cuni.cz - http://www.ff.cuni.cz/departments/fu/ip-intro.html

ABSTRACT

The natural speech communication is necessarily always completed by a prosodic realisation of the utterance. The bad identification of the marked prosody causes many communication problems, especially when the speaker and the listener don't share the mother tongue. This article presents the partial results of the study comparing Czech and French marked prosodic realisations. Four groups of listeners had to identify the attitudes expressed in Czech and French utterances realised by both Czech and French speakers. It seems that the production and the perception of the prosody is conduced by the prosodic stereotypes of the mother tongue. This adaptation of prosodic schemas brings many important communication difficulties.

1. INTRODUCTION

Plus de 90% de la communication est réalisé verbalement, par l'émission des sons. La réalisation sonore contient obligatoirement une composante prosodique (suprasegmentale) du message. Ce niveau de message est également touché par l'idiolecte du locuteur. Une mauvaise compréhension et un mauvais décodage de la prosodie du discours ne sont pas, dans la communication quotidienne, des phénomènes rares. Combien de fois dit-on «Ce n'est pas **ce** que tu dis, mais **la façon** dont tu me le dis!» Et tout cela dans la situation où nous parlons la langue maternelle. La communication des locuteurs non-natifs est encore moins sûre au niveau de la compréhension de la prosodie.

Le locuteur communique, à l'aide de la prosodie, une information qui complète ou modifie le sens du discours. Cette situation est connue de la pragmatique et des actes de langage.

2. COMMUNICATION DES LOCUTEURS NON-NATIFS

Un petit exemple de mauvaise interprétation de la prosodie dans la communication des locuteurs qui ne partagent pas la langue maternelle: les Tchèques expriment leur accord ou bien assurent l'interlocuteur de leur attention par l' interjection «hm». Dans la communication avec un francophone, l'utilisation «tchèque» (affirmation) peut introduire un malentendu. Un francophone répétera ce qu'il venait de dire, parce que l'intonation de l'interjection utilisée n'est pas nettement descendante et il interprétera le «hm» comme une question et une demande de répétition. Pour qu'un francophone identifie bien l'affirmation exprimée, il faut utiliser l'intonation nettement descendante. Une telle intonation correspond, en tchèque, à une attitude spécifique - «incrédulité». Le changement prosodique qui représente une nuance phonostylistique en tchèque a, en français, une valeur phonologique.

3. EMOTIONS OU ATTITUDES?

Dans le domaine des expressions prosodiques des émotions et attitudes, la terminologie n'est pas tout à fait stabilisée. De plus, les termes *émotion* et *attitudes* n'ont pas le même sens pour les psychologues et pour les phonéticiens.

3.1 Emotions

Dans la présente étude, nous avons adopté le terme *attitudes* et non *émotions*. Les études sur les émotions sont orientées surtout vers la psychologie et la psychiatrie [Sche89] et aident à diagnostiquer des troubles psychiques ainsi qu'à évaluer l'efficacité du traitement chez les dépressifs, schizophréniques, etc. On parle des expressions émotives dans la parole lorsqu'il s'agit de reflets spontanés des changements de l'état affectif et/ou intellectuel. Les expressions acoustiques des émotions sont conditionnées par des changements physiologiques. Rythme cardiaque, rythme respiratoire, tension musculaire (surtout dans le larynx) sont les phénomènes qui influencent le débit, l'intensité et le F0 de la parole et le timbre de la voix.

Les conditions physiologiques des changements acoustiques étant données, on peut supposer que les expressions prosodiques des émotions sont universelles, communes pour toutes les langues. Le locuteur ne contrôle pas entièrement la réalisation prosodique du discours. Dans le modèle de Buehler, il s'agit de l'expression du symptôme (la prosodie devient le symptôme de l'état psychique du locuteur). La fonction expressive domine dans le discours [Sche88].

3.2 Attitudes

La situation est différente dans l'expression des attitudes: tout d'abord, il s'agit de l'expression volontaire de l'attitude du locuteur. Deuxièmement, les expressions attitudinales sont orientées vers l'auditeur [GaGui94]. La réalisation prosodique n'est pas conditionnée physiologiquement, du moins pas directement. Elle peut être motivée par l'expression prosodique de l'émotion correspondante, mais elle varie aussi d'une langue à

l'autre conformément au système prosodique de la langue concrète.

4. Production et perceptions de la prosodie marquée

Dans une langue étrangère, la prosodie est perçue et produite selon certains schémas: le locuteur introduit les schémas prosodique de sa langue maternelle qui peuvent représenter, pour la langue utilisée, un élément étranger et en tant que tel rendre la communication plus difficile. La prosodie étrangère est interprétée selon la signification des schémas prosodiques ressemblants dans la langue maternelle [Mor69], [CrFe83].

Tableau 1: Résultats des tests perceptifs: identification correcte en pourcent.

groupe d'auditeurs		**partie du test**	
		française	*tchèque*
Français	a)	**34%**	**31%**
	b)		**37%**
Tchèques	c)	**51%**	**59%**
	d)	**54%**	

5. Expérience

Le présent article est centré sur les attitudes et les types de réalisations prosodiques qui causent des problèmes dans la communications des locuteurs avec la langue maternelle tchèque et française. Deux corpus parallèles ont été créés, tchèque et français. Chacun contient 5 différentes phrases affirmatives, 5 questions totales et 5 questions partielles. 5 locuteur tchèques et 2 locuteurs français (tous bilingues) ont eu pour tâche de réaliser toutes les phrases dans les deux langues en exprimant 8 attitudes suivantes: 1. neutralité (comme référence), 2. joie, 3. admiration, 4. surprise, 5. ennui, 6. tristesse, 7. colère, 8. peur. Les phrases étaient sémantiquement neutres et ne variaient pas selon l'attitude. Ceci est une des méthodes utilisées dans les études comparables [AutCri72], [BezBov84], [GaGui94], [Moz98]).

5.1 Tests auditifs

Les phrases ainsi obtenues ont été testées par 4 groupes d'auditeurs:

a) les Français qui ne comprennent pas le tchèque;
b) les Français qui comprennent le tchèque;
c) les Tchèques qui ne comprennent pas le français;
d) les Tchèques qui comprennent le français.

Les auditeurs étaient étudiants en philologie à la Faculté des Lettres de l'Université Charles, en phonétique à l'Université Paris 7 et les étudiants de tchèque à Paris 4.

La différenciations des groupes a et b, c et d permet de définir la mesure dans laquelle la compréhension de la langue utilisée facilite l'identification correcte de l'attitude exprimée par la prosodie.

Les résultats des tests perceptifs sont résumés dans le tableau 1.

6. Hypothèses proposées

Les résultats de l'analyse statistique ont permis de proposer et de vérifier trois hypothèses:

1. La connaissance de la langue utilisée facilite l'identification des attitudes transmises par la prosodie: **OUI** (**69%**)
2. La réalisation des énoncés par un locuteur natif facilite l'identification des attitudes transmises par la prosodie: **NON** (**38%**)
3. Le fait que les locuteurs et les auditeurs partagent la langue maternelle facilite l'identification des attitudes transmises par la prosodie: **OUI** (**62%**)

Les hypothèses coroborrées ont orienté notre attention vers certains types de schémas prosodiques ainsi qu'à l'expression prosodique de certaines attitudes, intéressantes du point de vue de la différence dans la perception de la langue maternelle et étrangère.

7. Fautes d'identification des attitudes chez les auditeurs tchèques et français

Dans la partie suivante, quelques fautes d'identification par les auditeurs tchèques et français respectivement seront analysées: quant à la durée et à l'intensité, il est très intéressant de comparer les réalisations prosodiques des énoncés 1 et 2. Les auteurs des études phonostylistiques ([Fón82], [Lé92]) soulignent surtout l'importance de l'intonation; selon eux, ce sont les changements de la courbe mélodique qui provoquent les nuance du sens perçues.

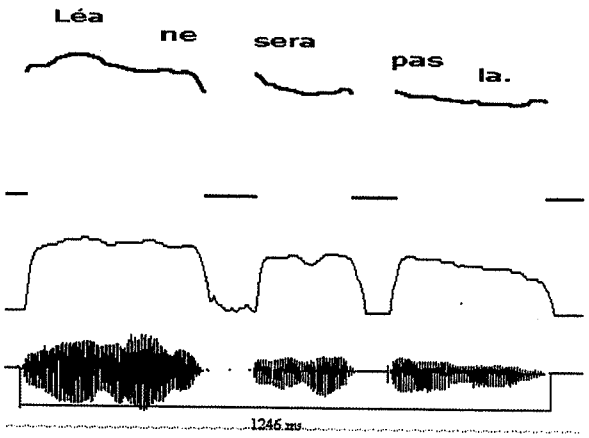


Figure1: Un énoncé français «ennuyé» identifié comme «triste» (48-1.wav).

7.1 Regroupement triste - furieux

Les deux premiers énoncés présentés ici ont été réalisés par une locutrice française, avec l'intention d'exprimer l'ennui et ils ont été bien identifiés par les auditeurs français. Les significations des deux énoncés (de leur prosodie plus exactement) ont été différentes pour les auditeurs tchèques: bien que les courbes mélodiques des énoncé soient pratiquement identiques, le premier a été identifié comme «triste» et le deuxième comme «furieux».

Valeurs de la durée et de l' intensité Il est étonant que l'écart de F0 (différence entre F0 minimal et maximal) soit plus important dans le premier énoncé (81 Hz) qui a été identifié comme «triste». La fréquence fondamentale varie entre 164 et 245 Hz. L'écart mélodique du deuxième énoncé, identifié comme «furieux», est de 68 Hz, avec les variations mélodiques entre 171 et 239 Hz. La différence des valeurs d'écart est relativement petite, mais rappelons que la tristesse est le facteur qui, d'après la plupart des auteurs, minimalise la variabilité mélodique de la parole. La différence entre les deux énoncés est manifestement présente dans la durée: le premier, «triste», a été réalisé pendant 1246 ms (quelque 5 syllabes par seconde), le deuxième, «furieux», pendant 639 ms (9 syllabes par seconde environ). En plus, un sommet énergique saillit sur la première syllabe de l'énoncé «furieux».

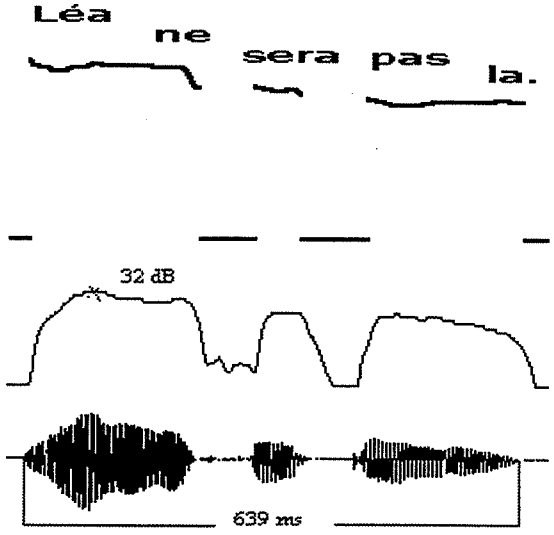


Figure2: Un énoncé français «ennuyé» identifié comme «furieux» (48-2.wav).

Nous pouvons donc affirmer que l'intonation n'est pas, malgré les opinions adoptées, le seul phénomène qui influence l'identification de l'attitude et que la durée et l'intensité contribuent d'une façon remarquable à la perception des nuances dans le sémantisme prosodique.

7.2 Regroupement joyeux - étonné

La deuxième paire d'énoncés a été réalisée par deux locuteurs tchèques; le premier comme «neutre»; il a été identifié correctement par les auditeurs tchèques. La réalisation prosodique a orienté l'identification par les auditeurs français vers la catégorie «joyeux». L'analyse des énoncé français identifiés de cette manière mène à la remarque qu'une prosodie «joyeuse» française est caractérisée par un timbre clair, une haute mélodicité (les différences importantes de F0 des syllabes successives), par les variations de l'intensité et du rythme (forte accentuation des syllabes accentuées; l'expression prosodique de la tristesse, par exemple, est caractérisée par une mélodicité très peu variée et par les différences énergiques faibles entre les syllabes accentuées et non accentuées). Les différences de F0 entre les syllabes successives sont, dans le troisième énoncé, de 60, 10 et 100 Hz. Le débit est relativement rapide (955 ms, soit 6 syllabes par seconde).

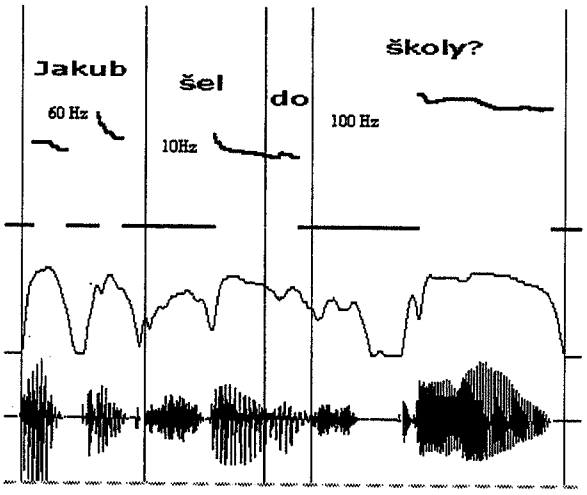


Figure3: Un énoncé tchèque «neutre» identifié comme «joyeux» (48-3.wav).

Quels changements prosodiques ont mené les auditeurs français à identifier cet énoncé comme «joyeux»? La clarté du timbre, le niveau relativement élevé de F0, la grande mélodicité de l'énoncé et le rythme marquant accompagnent d'habitude la prosodie «joyeuse» française. La comparaison avec la courbe d'intensité du premier énoncé (identifié comme «triste») soulignera l'effet de l'accentuation régulière.

L'énoncé représenté par la figure 4 a été également identifié par les auditeurs français comme «joyeux». Sa courbe mélodique partage certains formes avec celle de l'énoncé no 3. Le niveau de F0 est élevé, la mélodicité grande et la différence entre la hauteur des syllabes successives assez importante. L'intensité est de même variabilité comme dans l'énoncé 3.

L'énoncé 4 a été réalisé par une locutrice tchèque avec l'intention d'exprimer la surprise et il a été bien identifié par les auditeurs tchèques. L'intonation «étonnée» tchèque est caractérisée par une différence importante de F0 entre deux syllabes successives – le plus souvent la première et la deuxième syllabe du dernier groupe rythmique. Dans cet énoncé, la différence atteint 170 Hz et devient ainsi le guide de l'identification de l'attitudes exprimée pour les auditeurs tchèques (la surprise dans ce

cas). Quelle devrait être la prosodie de l'énoncé pour qu'il soit identifié comme «joyeux» par les auditeurs tchèques?

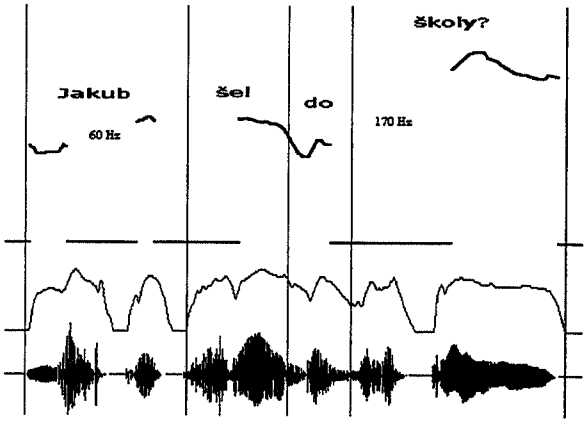


Figure 4: Un énoncé tchèque «étonné» identifié comme «joyeux» (48-4.wav).

Les analyses des énoncés réalisés et identifiés comme «joyeux» par les Tchèques montrent qu'une prosodie «joyeuse» nécessite un débit plus rapide et un timbre de la voix plus clair. Dans le spectre de l'énoncé mentioné ici, le bruit apparaît dans la partie centrale du spectre.

8. Conclusion

Dans le présent article, nous avons essayé de montrer quelques réalisations prosodiques de certaines attitudes qui peuvent être mal interprétées dans la communication des locuteurs non-natifs. Il est évident que la prosodie mal identifiée par les auditeurs ne partageant pas la langue maternelle avec les locuteurs est plutôt de caractère conventionnel et diffère d'une langue à l'autre. La prosodie attitudinale bien identifiée est motivée par la prosodie des émotions correspondantes (qui se montre plus universelle).

La motivation extralinguistique des expressions prosodiques des attitudes n'est que partielle et c'est ce qui la distingue des expressions prosodiques de l'émotions qui sont conditionnées par des changements physiologiques. La prosodie attitudinale peut être considérée comme un type de signe. En tout cas, il s'agit d'un signe asymétrique: une attitude peut être exprimée par plusieurs manifestations prosodiques et vice versa, une réalisation prosodique peut renvoyer à plusieurs attitudes selon le contexte.

Il faut avouer qu'il n'est pas possible de recommander aux étudiants ou aux locuteurs d'une langue étrangère un modèle de la réalisation prosodique bien défini tout en les assurant que leurs interlocuteurs comprendront l'attitude exprimée. La perception de la langue, qu'elle soit maternelle ou étrangère, est un processus complexe; la parole est perçue comme un ensemble de divers paramètres et de relations entre eux. La communication des attitudes (production et perception) est influencée par le contexte: du contexte le plus proche (verbal et nonverbal) dans le discours – parce que la prosodie devient marquée au moment où il y a matière à comparer - , jusqu'au large contexte de la personnalité du locuteur et de ses expériences.

Bibliographie

[AutCri72] Autesserre, D., DiCristo, A. (1972) Recherches psychosémantiques sur l'intonation de la phrase française, Travaux de l'Institut de Phonétique d'Aix, Aix en Provence

[BezBov83] Bezooijen, R. van, Boves, L. (1983) The Relative Importance of Vocal Speech Parameters for the Discrimination of Prosody, Proceedings of the Xth ICPhS, Utrecht

[CrFe83] Cruz-Ferreira M. (1984) Perception and Interpretation of Non-Native Intonation Patterns Proceedings of the Xth ICPhS, Utrecht

[Fón82] Fónagy, I. (1982) La vive voix, Payot, Paris

[GaGui94] Galazzi, E., Guimbretière, E. (1994) Intonation et attitudes: une question de perception, in: Studi di Linguistica, Storia della lingua Filologia francesi, Edizioni dell'Orso, Milan

[Lé92] Léon, P. (1992) Précis de phonostylistique. Parole et expressivité, Nathan, Paris

[Mor69] Morávek, M. (1969) Lidská řeč, Orbis, Praha

[Moz98] Mozziconacci, S.J.L. (1998) Speech Variability and Emotion: Production and Perception, Technische Universiteit Eindhoven

[Sche88] Scherer, K. (1988) On the Symbolic Function of Vocal Affect Expression, Journal of Language and Psychology, 2, pp. 79 – 100

[Sche89] Scherer, K. (1989b) Vocal Correlates of Emotional Arousal and Affective Disturbance, Handbook of Social Psychology, H.Wagner and A.Manstead Eds., John Wiley & Sons Ltd.

Les sosies vocaliques
Inversion et focalisation

L.J. Boë, C. Abry, D. Beautemps, J.L. Schwartz, R. Laboissière

Institut de la Communication Parlée - CNRS / INPG / Université Stendhal
Domaine Universitaire BP 25 38040 Grenoble Cedex 9
Tél.: ++33 (0)476 82 43 38 - Fax: ++33 (0)476 82 43 37
Mél: boe@icp.inpg.fr - http://www.icp.inpg.fr

ABSTRACT

A realistic articulatory model generates vowels, identic at the acoustic level but different in terms of phonetic classication. Are these "vowel doubles" realized by speakers, and in which conditions? The questions are discussed in the general frame of inversion of articulatory models and in the field of the theory of focalization.

1. LE CADRE DE L'INVERSION

En inscrivant leur recherche dans le cadre théorique général de *l'inversion* articulatori-acoustique, Atal & al. [Ata78] ont bien montré que pour un système possédant n commandes d'entrée et produisant une sortie dans un espace à p-dimensions, avec n > p il existe des courbes – les fibres (*fibers*) ou feuillets (*manifolds*) – dans un espace R^{n-p} le long desquelles les variations des commandes ne produisent pas de modifications en sortie. Cette propriété des systèmes, dits *surnuméraires*, bien connue en robotique, pose de sérieux problèmes pour l'inversion puisque, sans contrainte, il existe tout un ensemble de commande possibles qui correspondent à un même signal de sortie.

Même intrinsèquement « bien contraint » un modèle articulatoire peut donc générer, s'il est surnuméraire (ce qui est généralement le cas), des voyelles identiques au niveau acoustique mais présentant des dispositions articulatoires différentes en termes de classification phonétique. Ces configurations – que nous appellerons *sosies vocaliques* – sont-elles effectivement réalisées par les locuteurs et dans quelles conditions ? Nous tenterons de répondre à ces questions en les situant dans le cadre général de *l'inversion* des modèles articulatoires et en réinterprétant les résultats dans le champ de la *Théorie de la focalisation* [Abr89] [Bad89] [Sch97].

2. STRATEGIE ET CRITERES D'ATTEINTE DE CIBLE VOCALIQUES

Pour dresser une typologie des sosies vocaliques, nous avons choisi un modèle articulatoire réaliste [Mae89] intégrant l'essentiel des contraintes physiologiques relatives au jeu des lèvres, aux mouvements de la mâchoire et de la langue et à la position du larynx Les articulateurs sont clairement identifiables avec 7 degrés de liberté : 2 pour les lèvres (la hauteur *LipH* et la protrusion *LipP*), 1 pour la position de la mâchoire (*Jaw*), 3 pour le contrôle du corps, du dos et de la pointe de la langue (*Tongue Body*, *Tongue Dorsum* et *Apex*) et 1 pour la position du larynx (*Lx*).

Nous avons sélectionné 3 classes de voyelles qui couvrent assez bien l'espace vocalique : 7 voyelles périphériques [i e ɛ a o ɔ u], 3 voyelles antérieures labalisées [y ø œ] et la voyelle haute [ɯ]. Nous disposons, pour ces cibles, de valeurs de commandes articulatoires et acoustiques prototypiques, ajustées par expertise phonétique par rapport à l'espace acoustique maximal du modèle et validées perceptivement [Val95] (Tableau 1).

V	F_1	F_2	F_3	F_4
i	225	2245	3355	4180
e	400	2040	2620	3385
ɛ	630	1590	2310	3335
a	770	1230	2290	3235
o	430	800	2145	3405
ɔ	585	880	2220	3440
u	275	690	2195	3485
y	230	1880	2110	3085
ɯ	300	1205	2200	3415
ø	410	1740	2155	3165
œ	545	1555	2180	3240

Tableau 1. Les valeurs formantiques des prototypes vocaliques.

Pour atteindre une cible acoustique à partir d'une configuration articulatoire initiale, la procédure d'inversion que nous avons adoptée ici a consisté, à chaque itération, à générer au hasard un ensemble de N jeux de paramètres de commande articulatoire, distribués uniformément dans un intervalle ± delta et ayant comme moyenne la position courante. Le premier jeu de paramètres qui permet de se rapprocher de la cible, au sens d'une distance calculée à partir de F_1 F_2 F_3 (en Hz ou en bark), étant pris comme nouvelle valeur courante et ainsi de suite jusqu'à atteinte de la cible. Pour améliorer la vitesse de convergence, la valeur delta (ajustée initialement en fonction du maximum maximorum des distances acoustiques, soit la distabce entre [i] et [a]) est diminuée en fonction inverse du logarithme du nombre cumulé des tirages.

La figure 1 présente, dans le plan F_1 F_2, une atteinte de cible pour les 11 voyelles choisies, à partir d'une position initiale correspondant à la valeur moyenne des paramètres articulatoires du modèle (p = 0).

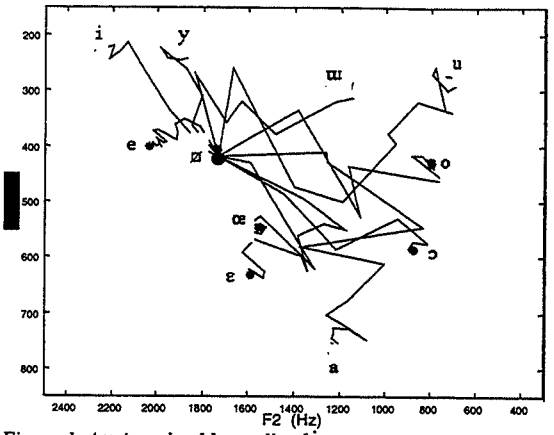


Figure 1. Atteinte des 11 voyelles [i e ɛ a y ø œ ɯ u o ɔ] à partir de la position neutre du modèle articulatoire.

Pour parcourir des chemins articulatoires différents, et espérer ainsi produire le maximum de sosies (variantes contextuelles), nous avons effectué 100 atteintes de cible pour chaque voyelle. La figure 2 montre que l'espace maximal est effectivement bien couvert par ces 1100 parcours. À titre indicatif, dans l'environnement informatique utilisé, chaque cible est atteinte en moins de 10 secondes.

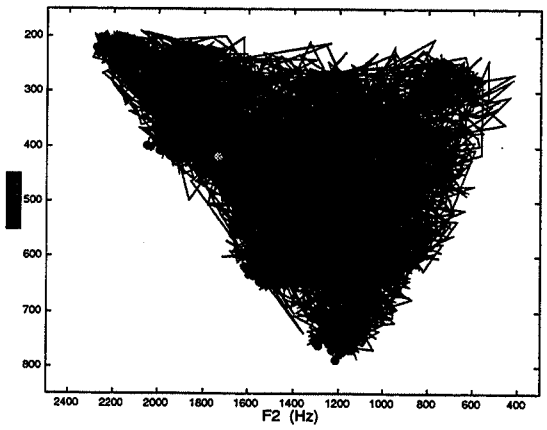


Figure 2. L'ensemble des atteintes de cible pour les 100 trajets.

Nous avons choisi plusieurs critères d'atteinte de cible associés à différentes hypothèses perceptives :

• *Critère 1 :* une simple contrainte acoustique ; on considère que la cible est atteinte si la distance du point courant à la cible est inférieure aux seuils différentiels établis perceptivement [Fla55], soit 5% pour F_1 et 10% pour F_2 F_3. Par précaution nous avons retenu 3.5% pour F_1 et 7.5% pour F_2 et F_3.

• *Critère 2* : une contrainte perceptive fondée sur F'_2 (évalué en bark) [Sch89] ; nous avons retenu une distance déjà validée pour la prédiction des systèmes vocaliques [Sch97] :
$d = ((F_1 - F_{1cible})^2 - 0.09\,(F'_2 - F'_{2cible})^2)^{1/2}$
avec un seuil d'atteinte correspondant à des différences de quelques % sur les 3 premiers formants.

• *Critère 3 :* une spécification dans le cadre de la théorie de la focalisation selon laquelle l'objectif est de regrouper deux formants successifs autour d'un point focal – perceptivement prégnant – qui caractérise la voyelle, soit pour le modèle (un conduit vocal d'homme) (Tableau 2) :

Voyelle	Focalisation	Point focal (Hz)	Condition nécessaire et suffisante
[i]	F_3F_4	3700	F_3 max
[y]	F_2F_3	2000	F_2F_3 focalisés + aire aux lèvres minimale
[a]	F_1F_2	1000	F_1 max
[u]	F_1F_2	500	F_2 min

Tableau 2 Focalisation, conditions nécessaires et suffisantes.

Avec les deux premiers critères, les résultats sont identiques (figure 3). On retrouve bien, pour chaque voyelle la configuration qui sert de prototype dans toutes les descriptions phonétiques. Les différences permettent de retenir 5 classes vocaliques :

• Les trois voyelles *ouvertes* [ɛ a œ] se caractérisent par une très grande stabilité articulatoire des lèvres et de la langue. On ne peut qu'être frappé par la très faible plasticité du [a] : l'analyse des paramètres articulatoires des simulations montre qu'un recul–abaissement de la langue est systématiquement requis pour assurer la constriction pharyngale, alors que la position mâchoire, pourvu qu'elle soit légèrement abaissée, n'est pas déterminante.

• Nettement plus variables, [i e] présentent des différences d'aperture labiale compensables par des déplacements du corps de la langue : une plus grande ouverture associée à un recul de langue. On ne peut pas dire que ces 2 voyelles possèdent deux sosies, mais elles présentent nettement des possibilités allophoniques contextuelles. La grande variabilité de l'aire aux lèvres de [i] a déjà été notée : [si] vs. [ʃi] [Abr86].

• Des sosies vocaliques pour les voyelles [y ø] qui peuvent être réalisés [+ LAB] (prototype) ou [– LAB], c'est-à-dire – protrus, cette disposition des lèvres étant compensée par un recul (ou abaissement) de la langue, comme attesté dans le contexte ouvrant [ʃ] [Abr86].

• Deux sosies vocaliques pour [ɯ] : avec un abaissement du dos pour compenser une fermeture non-prototypique. Il s'agit, à notre connaissance, d'une première mise en évidence.

• Trois soisies pour [u o ɔ] : vélo palatal, vélo-pharyngal et pharyngal. Pour [u o] il avait été prédit [Boë92], [Boë96] les deux premiers lieux d'articulation. Émerge ici un troisième sosie, nettement pharyngal, avec un larynx abaissé. Toutes lès positions intermédiaires correspondent à un déplacement le long de la fibre. Ces trois sosies se classifient bien en fonction de valeur du paramètre dos de la langue :

$$Drsm > 2, \quad 2 \leq Drsm \leq -1 \quad \text{et} \quad Drsm < -1$$

Il est important de noter que les trois sosies du [u] sont tous protrus et fermés. À première vue, ce résultat ne semble pas confirmer les travaux de [Sav95] dans le protocole de production de [u] avec la contrainte du *lip-tube*. En fait, il ne s'agit pas d'une contradiction : avec les seuils que nous nous sommes fixés, tant au niveau articulatoire (constriction minimale de 0.1 cm^2) qu'acoustique (distance au prototype), il n'est pas

possible de produire un [u] lèvres ouvertes Mais, avec plus de latitude (constriction minimale à 0.05 cm^2), on obtient bien, avec notre algorithme d'inversion, un [u], lèvres ouvertes à 3 cm^2. Avec des différences sur F_1 et F_2 de l'ordre de 16 % (et 8% sur F_3) par rapport à notre prototype, il présente un des types de configuration caractéristique observé par [Sav95]. On retrouve la tendance de tous les sujets testés à postérioriser pour compenser l'ouverture labiale. Pour certains sujets les décalages en fréquence (de F_2) tendant à être compensés par un Fo élevé (effet Traunmuller, [Sav99]).

Le cas du critère n° 3. Nous retrouvons très précisément les prototypes [i a u], avec le critère de la focalisation et même en utilisant les implications (=>) acoustiques liées aux contraintes du conduit vocal. Pour les trois voyelles cardinales, les résultats sont atteints avec :

F_3 max => F_1 min F_2 max F_4 proche de F_3 => [i]
F_1 max => F_2 à 1200 Hz et F_3 à 2300 Hz => [a]
F_2 min => F_1 min , F_3 autour de 2200 Hz => [u]

Mais la focalisation F_2 F_3 n'implique pas F_1. Sans contrainte aux lèvres émerge un [y] suédois (focalisation à 2150 Hz, [– LAB] (lèvres ouvertes) proche du [i] et non du [y] du français, qui est un peu moins bien focalisé (à 2000 Hz) et [+ LAB] (lèvres fermées).

3. Un premier bilan

Cette typologie des sosies, réalisée dans le cadre de l'*inversion* d'un *modèle articulatoire réaliste,* ne prétend pas à l'exhaustivité. Des configurations attestées ici émergent plusieurs classes de voyelles selon les possibilités compensatoires. Les stratégies mises en évidence combinent pour l'essentiel un jeu des lèvres (aperture, protrusion) à un déplacement de la langue (antéro-postérieur) ou à son abaissement. On peut noter, au passage, qu'aucune d'entre elles ne confirme la *propriété d'anti-symétrie*, résultat pourtant essentiel, des *méthodes variationnelles* appliquées aux *fonctions d'aire*. Le conduit vocal ne se comporte pas comme un jeu de n-tubes indifférenciés.

Qu'est-il est possible par *inversion* d'un modèle articulatoire *réaliste* d'inférer du son sur l'articulation :

- [ɛ œ a] : la position des lèvres et la forme intégrale de la langue ;
- [y ø] : des lèvres ± labialisées et une articulation pré ou post-palatale, donc pas de précision sans vision ;
- [ɯ] : une articulation palatale ou pharyngale associée à un dos + haut ou + bas ;
- [u o ɔ] : une constriction linguale postérieure (de palatale à pharyngale) ; pour [u], des lèvres fermées (< 0.8 cm^2), une position des lèvres ± fermée pour [o] et ± ouverte pour [ɔ].

Les ambiguïtés labiales correspondent bien à celles qui ont été mises en évidence :

- en production, par [Dja89] pour les voyelles antérieures [+ Lab] et les centrales ;
- en perception, par [Lis92] sur 18 qualités vocaliques (indécision le trait [± Round] de [ɯ] et de [ɔ] et de [œ] avec l'audio seul).

Enfin, et ce n'est pas le moindre résultat, la focalisation se révèle un fil conducteur opératoire pour expliquer les paramètres acoustiques cruciaux pour les stratégies articulatoires des voyelles focales : F_3 pour [i], F_2 pour [u], F_1 pour [a], F_2F_3 pour [y].

References

[Abr96] Abry, C, Boë L.-J. (1986) "Laws for lips", *Speech Comm.* 5, 97-104.

[Ata78] Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W. (1978) "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique", J. Acoust. Soc. Am., Vol. 63, pp. 1535-1555

[Bai 91] Bailly, G., Laboissiere, R., Schwartz, J.-L. (1991) "Formant trajectories as audible gestures: An alternative for speech synthesis", *J. Phonetics*, 19, 9--23.

[Bad90] Badin, P., Boë, L.J., Perrier, P., Abry, C. (1990) "Acoustic considerations upon formant convergence", *J. Acoust. Soc. Am.*, 63, 1535-1555.

[Boë92] Boë, L.J., Perrier, P., Bailly, G. (1992), "The geometric vocal tract variables controlled for vowel production : proposals for constraining acoustic-to-articulatory inversion", *J. Phonetics*, *20*, 27-38.

[Boë95] Boë, L.J., Gabioud, B., Perrier, P., Schwartz, J.L., & Vallée, N. (1995), "Vers une unification des espaces vocaliques", *Levels in Speech Communication: Relations and Interactions*, 63-71. Elsevier B.V.

[Boe96] Boë, L.J., Schwartz, J.L., Laboissière R., Vallée, N. (1996) *"Integrating articulatory-acoustic constraints in the prediction of sound structures"*, 1st ESCA Workshop on Speech Production Modelling, 163-166

[Fla55] Flanagan, J. (1955) "A difference limen for vowel formant frequency", *J. Acoust. Soc. Am.*, 27, 613-617.

[Lis92] Liker, L., Rossi, M. (1992) "Auditory and visual cueing of the [±rounded] feature of vowels", *Language and Speech*, 35, 391-417.

[Mae89] Maeda, S. (1989) *"Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model"*, In Speech Production and Modelling, 131-149 Academic Publishers, Kluwer.

[Sch89] Schwartz, J.L., Escudier, P. (1989) "A strong evidence for the existence of a large-scale integrated spectral representation in vowel perception", *Speech Communication*, 8, 235-259.

[Sav95] [Sav99] Savariaux, C., Perrier, P., Orliaguet, J.P. (1995), "Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production", *J. Acoust. Soc. Am.*, *98*, 2428-2442, "II. Perceptual analysis", *J. Acoust. Soc. Am.*, 106 , 381-393.

[Sch97] Schwartz, J.L., Boë, L.J., Vallée, N., & Abry , C. (1997), "The dispersion-focalization theory of vowel systems", *Journal of Phonetics, 25*, 255-286.

[Val95] Vallée N., Boë L.J., Payan J. (1995), "Vowel prototypes for UPSID's phonemes", XIIIth Int. Congr. Phonetic Sciences, 1, 424-427.

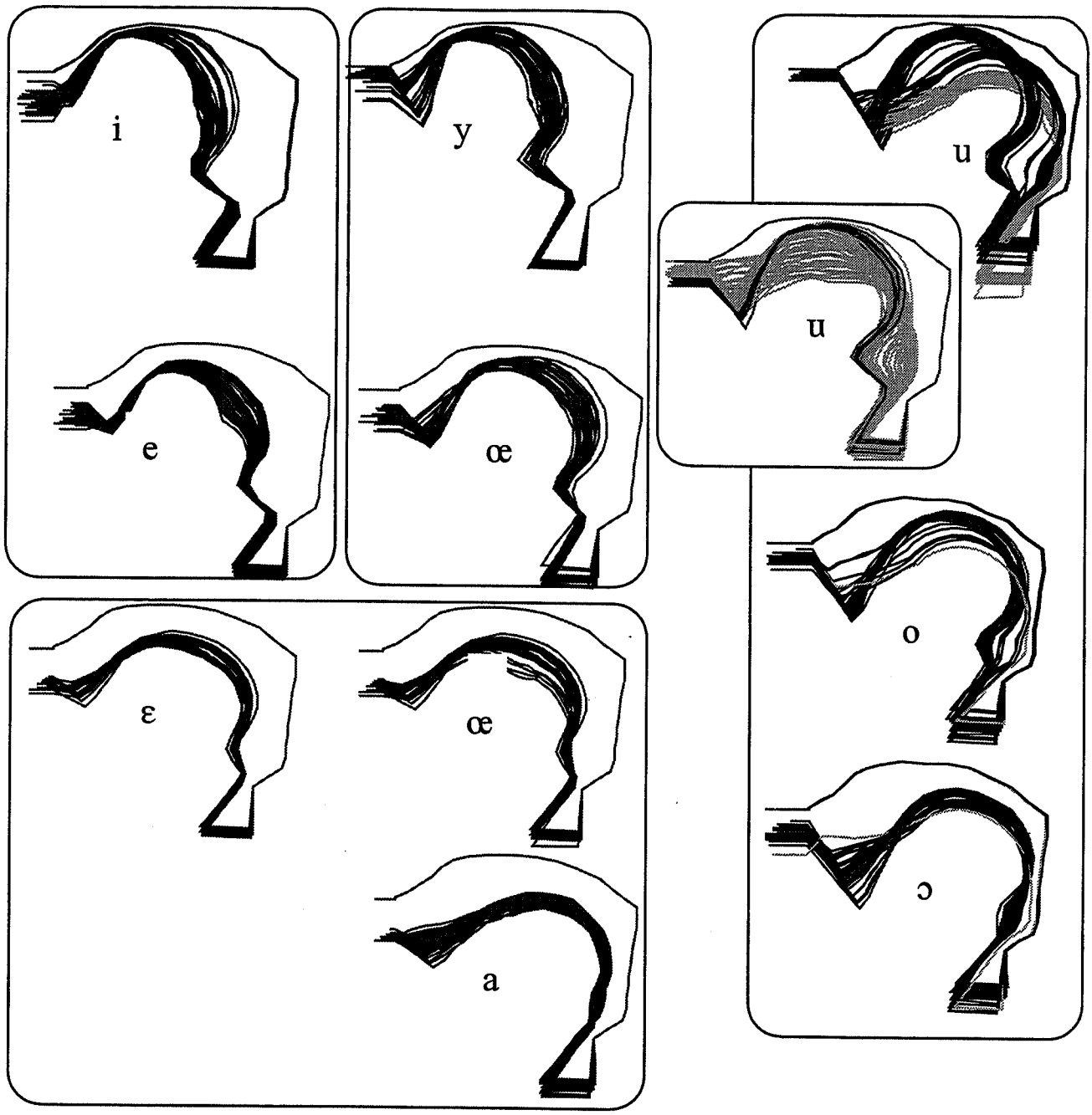


Figure 3. Les sosies vocaliques.

	i e	ɛ œ a	y ø	o ɔ	ɯ	u
Sosie1	Prototype	Prototype	Prototype	Contexte palatal	Prototype	Prototype (vélo palatal)
Sosie 2	X	X	En contexte [ʃ] [– LAB]	Prototype (vélaire)	Contexte pharyngal	Contexte vélaire (vélo pharyngal)
Sosie 3	X	X	X	Contexte pharyngal	X	Contexte pharyngal

Tableau 3. Le classement des sosies.

Vers une modélisation de la durée des sons pour la génération automatique du rythme dans la synthèse de la langue arabe

Z. Zemirli, N. Vigouroux

Institut de Recherche en Informatique de Toulouse
UMR CNRS 5505 – UPS - 118, Route de Narbonne 31062 Toulouse Cedex
Tél.: ++33 (0)561556314 - Fax: ++33 (0)561556258
Mél: {zemirli,vigourou}@irit.fr

ABSTRACT

This article describes our approach to modelize the sound duration of standard Arabic speech. The final aim is the prosodic generation for Arabic Text-to-Speech synthesis. Several authors have already identified on isolated word corpora the effects of speaker, word structure, sound nature, and Arabic dialects. With this front-end, our work hypothesis is to verify these effects for standard Arabic on **continuous speech**. Firstly a speech corpora of continuous speech was defined and recorded. Then, two automatic tools –the SYNTHAR+ grapheme-phoneme system and the MBROLIGN alignment – were used to label the speech data base recorded. Finally several contextually analysis of phonetic unit duration are conducted. Our studies on continuous speech confirm the results on isolated words –context effects on vowel duration, doubled consonants. We point out also the effect of the syllabic structure for the Arabic vowels.

1. INTRODUCTION

Les systèmes actuels de synthèse vocale à partir du texte pour les langues latines tels que le français, l'anglais, l'allemand, etc. sont capables de produire une parole synthétique de bonne qualité, naturelle et intelligible [Dut97]. Pour le français, cette qualité est le fruit de recherche fondamentale consacrée depuis des décennies à l'étude de la prosodie [Bar87], [Ros88], [Boe96], [Tou97], [DiC98], [Mal98], [Mer99] et [Nag98] pour ne citer que ceux-là. Cette qualité est due essentiellement à l'utilisation de techniques d'apprentissage automatique et d'outils d'étiquetage de la parole pour les traitements prosodiques [Boe96] ainsi que la disponibilité de grandes bases. [Bar87] a proposé un ensemble de règles qui détermine la durée segmentale en fonction de marqueurs syntaxico-prosodiques relatifs au mot, la position de la syllabe, la position du phonème dans la syllabe, etc.

Concernant la langue arabe, il n'existe à notre connaissance que deux systèmes commercialisés de synthèse à partir du texte : Multivox de la société d'AKADIMPEX et celui des sociétés Lernout& Hauspie et SAKHR-JVC. De nombreux travaux se sont intéressés à la variation de la durée des sons de la langue Arabe en fonction de différents facteurs : locuteurs, processus phonologiques, structures syllabiques, gémination et dialectes, etc. Parmi eux, Ghazali et Al [Gha92] ont étudié le traitement des processus phonologiques dans la synthèse à partir du texte par TD_PSOLA. Jomaa [Jom94] a traité de l'opposition de la durée vocalique en arabe tandis que Amrouche et Al [Amr98] se sont intéressés à la variation de la durée vocalique dans des structures syllabiques. Mentionnons que ces études reposaient essentiellement sur des corpus de mots isolés et de non mots.

Alors que l'étude de Zemirli [Zem98] souligne le rôle majeur de la durée des sons dans l'intelligibilité du système de synthèse Multivox, peu de travaux encore, sont consacrés à la génération du rythme et des contours mélodiques pour la synthèse de la langue arabe.

Des trois paramètres prosodiques –fréquence fondamentale, durée et intensité–, la durée des sons reste la plus difficile à modéliser. Celle-ci dépend du contexte de réalisation des phonèmes : nature, taille et structure de la syllabe, accent, etc. comme l'attestent les travaux référencés ci-dessus et ce, sur de la parole isolée. Qu'en serait-il pour des observations de parole continue utiles à l'élaboration de stratégies stylistiques sur le plan de durée pour la nouvelle génération des systèmes de synthèse de parole ?

Cette article présente la démarche méthodologique que nous avons adoptée pour la modélisation de la durée des sons en vue de la génération automatique du rythme dans la synthèse de la langue arabe standard.

Ce travail vise à identifier les effets du contexte immédiat sur la durée des phonèmes, l'allongement des consonnes géminées ainsi que l'influence de la structure syllabique sur la durée en parole continue. Pour conduire cette étude, nous avons procédé à l'élaboration d'un corpus d'étude de parole continue, représentatif de toutes les contraintes phonotactiques de la langue arabe. Nous décrirons ensuite succinctement les deux outils automatiques utilisés pour effectuer l'alignement des unités phonétiques sur le signal de parole. Des critères d'extraction des durées des unités phonétiques sont définis afin de tenir compte de la structure syllabique, des caractéristiques phonétiques de la langue arabe. Ces critères reprennent pour l'essentiel, ceux décrits dans [Bar87] mais ils sont adaptés aux caractéristiques phonologiques et syllabiques de la langue arabe.

Différents résultats illustrant l'effet de la structure syllabique, de la gémination et la position des phonèmes sur les variations de durée sont ensuite commentés. Une mise en oeuvre et une première évaluation du modèle de génération automatique du

rythme est discutée. Enfin, des perspectives à ce travail sont évoquées.

2. CLASSIFICATION PHONETIQUE ET STRUCTURE SYLLABIQUE

L'objet de ce paragraphe est de décrire succinctement une classification phonétique de la langue arabe ainsi qu'une brève description de sa structure syllabique. La langue arabe est constituée de 28 consonnes et de six voyelles de base : 3 voyelles courtes auxquelles s'opposent 3 voyelles longues. Une des caractéristiques de la langue arabe par rapport aux langues indo-européennes est l'emphase qui caractérise 5 consonnes notées suivant le formalisme utilisé dans MBROLA [Dut97] : s. (ص) d. (ض) t. (ط) et z. (ظ) et q (ق). Toutes les consonnes peuvent être géminées. Ci-dessous la classification des phonèmes de la base de sons AR1 pour la langue arabe utilisée par MBROLA (www.tcts.fpms.ac.be/synthesis). Les phonèmes /d./, /t./, /s./ et /z./ n'ont pas d'équivalent dans la notation SAMPA.

- Voyelles courtes : /a/ (ـَ), /i/ (ـِ) et /u/ (ـُ).
- Voyelles longues : /aa/ (ا), /ii/ (ي) et /uu/ (و).
- Voyelles emphatisées a. i. et u (après /d./ /t./ /z./ et /s./).
- Consonnes : /b/ (ب), /f/ (ف), /m/ (م), /w/ (و), /s/ (س), /z/ (ز), /n/ (ن), /r/ (ر), /l/ (ل), /s./ (ص), /S/ (ش), /Z/ (ج), /j/ (ي), /k/ (ك), /x/ (خ), /G/ (غ), /?/ (ء أ إ), /h/ (هـ), /X/ (ح), /H/ (ع), /t/ (ت), /d/ (د), /T/ (ث), /D/ (ذ), /t./ (ي), /d./ (ض), /z./ (ظ), et /q/ (ق).

La syllabe en langue arabe obéit à deux règles :

1. Le noyau syllabique est une voyelle.
2. Deux consonnes ne peuvent se suivrent sauf en fin de mot et devant une pause.

Ce qui permet d'obtenir les combinaisons syllabiques suivantes : CV **ka**taba , CVV **maa**ta, CVC **mak**tab, CVVC sa**laam**, CVCC **bint**. Une voyelle en début n'étant pas admise, elle est considérée comme la réalisation de la consonne hamza suivie d'une voyelle. Le système accentuel a, quand à lui, fait l'objet de nombreux travaux [Bla76], [Lec75] et [Zak84]. Bien que la prise en compte de l'accent soit nécessaire pour la modélisation de la durée, elle fera l'objet d'une prochaine étude.

3. DESCRIPTION DE L'APPROCHE METHODOLOGIQUE

3.1 Description des corpus acquis

Le corpus est composé de 150 phrases de parole continue représentant plus de 6000 phonèmes. Elles sont composées de deux à quinze mots. Elles ont été prononcées par deux locuteurs masculins arabophones, sans accent dialectal, selon trois consignes d'élocution lente (<10 phonèmes/sec), moyenne (10 à 13) ou rapide (>13). Elles représentent trois types de schémas intonatifs : affirmatif, interrogatif ou exclamatif. Le corpus a été acquis à 16 Khz et épuré des séquences de silence long de début de parole ou entre les mots.

3.2 Transcription des corpus

L'étape suivante a consisté à produire la chaîne phonétique au moyen du système SYNTHAR+ [Zem98]. Ce système assure la conversion des graphèmes arabes en phonèmes. L'entrée de SYNTHAR+ est une chaîne de graphèmes arabes voyellisés. La sortie est la chaîne phonétique compatible avec les entrées de MBROLA. Les performances de ce système ont été optimales : nous n'avons pas recensé d'erreur de transcription. Mentionnons toutefois que le corpus ne contenait pas d'abréviation, ni de sigle.

Cette transcription phonétique produite, nous procédons ensuite à l'alignement de celle-ci sur le signal au moyen de l'outil MBROLIGN disponible gratuitement sur le site www.tcts.fpms.ac.be/synthesis. L'alignement s'est avéré correct pour toutes les phrases et ce, pour les modes d'élocution lent et normal. Par contre, nous avons dû procéder à des réajustements des frontières de près de 5% des phonèmes pour toutes les phrases longues (plus de sept mots) en mode d'élocution rapide.

3.3 Génération des durées des unités phonétiques

Rappelons que la notion de longueur des mots de la langue arabe est très différente de la langue française. Celle ci se caractérise par un fort taux d'agglutination d'entités lexicales de base composant un mot. Par exemple, un mot peut être l'équivalent de toute une phrase : «أَرَيْتَهُم» qui signifie « les as tu vu ».

Après ces précisions, décrivons la procédure de calcul de la durée. Une fonction de MBROLIGN permet de générer un fichier résultat où à chaque phonème (y compris la pause) est associé son libellé, sa durée et éventuellement un ensemble de couples de nombres représentant la position du pitch en pourcentage par rapport à la durée totale du phonème et la valeur du picth à cette position. L'exemple ci-dessous donne un résultat pour la portion de phrase «هَلْ تَأْمُرُ», qui signifie « ordonnez-vous ».

_ **100**
h **80** 0 137 100 137
a **60** 16 142
l **50** 60 142
t **50** 20 142 60 142
a **70** 20 138
? **50**
m **50** 60 111
u **50** 20 105 80 102
r **50** 40 100 100 95

A partir de cette structure de données, nous avons élaboré une base de durée de l'ensemble des phonèmes du corpus. A chacun d'eux est associé, ses contextes droit et gauche en classe phonétique, des informations

syllabiques (position et nombre de phonèmes dans la syllabe). Celle-ci pourra être complétée ultérieurement par des informations accentuelles pour la poursuite de nos travaux.

4. ETUDE DE LA DUREE DES PHONEMES

4.1 Extraction

Nous avons procédé à plusieurs types d'extraction selon des critères de contextualisation afin :

1. d'étudier l'effet du contexte immédiat gauche et droit sur la durée du phonème ;
2. de mesurer la durée de l'allongement des phonèmes géminés ;
3. d'étudier l'influence de la position des phonèmes dans la syllabe sur sa durée ;
4. d'établir un rapport entre la taille de la syllabe et la durée des phonèmes.

4.2 Interprétation brute

Nous avons dressé le tableau 1 de résultats dit bruts qui comprend la durée ***moyenne minimale*** des phonèmes (Colonne 2) calculée sur la base d'une ***durée inférieure à 100 ms*** pour les phonèmes non géminés et la durée moyenne maximale (colonne 3) des phonèmes qui peut correspondre à un phonème géminé, à une voyelle longue ou un allongement particulier d'un phonème. A partir de ce tableau nous avons vérifié l'adéquation entre le phénomène de gémination (informations données par SYNTHAR+) et la durée des phonèmes issue de MBROLIGN. Il en ressort que les consonnes géminées voient le doublement de leur durée au minimum. Des tests de perception primaire ont confirmé la nécessité d'avoir cet allongement. Lors de mauvaises segmentations des consonnes géminées on ne perçoit plus la gémination à l'audition.

Tableau 1: Durées moyennes des phonèmes du corpus

Phonème	Durée moy min	Durée moy max
a	95	*
i	85	*
l	64	132
u	94	*
n	69	172
aa	*	160
m	68	166
t	73	147
r	68	156
j	63	120
b	72	145
w	60	125
H	87	180
?	71	135
k	78	143
s	94	163
h	78	150
d	66	125
ii	*	200
f	70	135
q	80	151
X	97	159
Z	86	130
a.	74	165
D	72	150
s.	77	149
z	80	144
d.	80	160
G	82	160
i.	63	154
x	88	141
uu	*	160
z.	76	156
S	87	175
T	92	170
t.	85	130
u.	71	151

* La durée des voyelles brèves est toujours inférieure à 100 ms tandis que celle des voyelles longues est toujours supérieure à 100 ms ce qui explique la non apparition de leur valeur dans ces cas.

4.3 Interprétation contextualisée

Nous avons défini deux coefficients de réduction CR1 (resp. CR2) de la durée de base (Db) de la voyelle dans les syllabes de type CVC (exp. **mak**tab) (resp. CVCG, exp. **sabb**aqa où CG est une consonne géminée) par rapport à la durée de la voyelle dans la syllabe de type CV. Le tableau 2 permet de constater que le coefficient de réduction CR2 est plus important que le coefficient de réduction CR1. La nouvelle durée Dv est égale à Db*(CR1 ou CR2) selon le type de syllabe. La durée des voyelles est donc plus réduite dans les syllabes de type CVCG que dans les syllabes de type CVC. Les valeurs de CR2 obtenues sur notre corpus confirment celles de [Amr98] qui mentionnait la réduction de la durée la voyelle devant une consonne géminée. Néanmoins ce coefficient de réduction est plus important pour des mots isolés que pour de la parole continue. Il faut également souligner que CR1 et CR2 dépendent de la nature de la voyelle.

Tableau 2: Durée moyenne des voyelles brèves suivant le type de syllabe.

	CV	CVC	CR1	CVCG	CR2
a	**97**	83	0.86	72	0.74
u	**98**	79	0.81	69	0.70
i	**87**	68	0.78	60	0.68

Le tableau 3 montre un allongement de la durée des phonèmes (en comparaison à celles du tableau 1) en fin de mot et devant une pause y compris les voyelles lorsqu'elle sont volontairement prononcées (habituellement la voyelle est omise à la fin d'un mot devant une pause). Cette augmentation est due à un effort fait par le locuteur pour la prononciation de la voyelle devant une pause. Le phonème **n** est un cas particulier, sa durée est plus que doublée. Après une étude plus fine sur ce phonème nous avons constaté que ce fort allongement est vérifié lorsqu'il marque le tanwin (marque d'indétermination). **CA** représente le coefficient d'allongement du phonème devant une

pause.

Tableau 3: Durée moyenne et coefficient d'allongement (par rapport à la durée de base du tableau 1) de quelques phonèmes en fin de mot et devant une pause.

Phonème	CA	Durée	Phonème	CA	Durée
a	1.25	120	**m**	**2.5**	**170**
u	1.22	115	**h**	**2.0**	**150**
i	1.29	110	q	1.37	110
s	**1.8**	**170**	**n**	**2.9**	**200**

Tableau 4: Facteur de réduction de la durée des phonèmes en fonction du nombre de syllabes du mot.

Nombre de syllabes	1	2	3	>3
CS	1	0.96	0.92	0.87
phonème l	70	67	64	60
phonème q	90	86	83	78
phonème m	72	69	66	62

Le tableau 4 illustre quelques exemples de relation entre le nombre de syllabes et la durée des phonèmes du mot. Plus le mot contient de syllabes, plus la durée des phonèmes est réduite sur l'ensemble du mot par le coefficient syllabique **CS**. CS est la moyenne de tous les CSi calculée par type de phonème.

5. MISE EN ŒUVRE ET EVALUATION

Afin d'évaluer les paramètres CS, CR et CA nous avons procédé à la génération automatique de la durée des sons sur un nouveau corpus nommé EVAL0 de 40 phrases prononcées par deux nouveux locuteurs arabophones et comportant quelques 1200 phonèmes selon les étapes suivantes :

1. Transcription automatique du corpus EVAL0 à l'aide de SYNTHAR+ et calcul des paramètres **CS** pour chaque mot et de **CRi** ou **CAi** pour les phonèmes (par défaut CS=CRi=CAi=1).
2. Génération automatique de la durée (Di) des sons à partir de la base (DMi) de durée moyenne établie, de CS, CRi et CAi. Di=DMi*CS*(CRi ou CAi).
3. Génération automatique de la durée des sons à l'aide de MBROLIGN.
4. Comparaison des durées générées par MBROLIGN avec celles produites par notre génération automatique en tenant compte des paramètres CS, CRi et CAi.

A l'issue de ces étapes nous avons constaté un écart moyen de 25 ms pour les voyelles et de 13 ms pour les consonnes. Bien que cet écart soit faible, nous procédons actuellement à une étude de l'effet de la ***durée des pauses*** sur la durée des phonèmes en début et fin de mot. Nous n'avions considéré que la pause à la fin des phrases dans cette étude.

6. CONCLUSION ET PERSPECTIVES

A l'issue de cette étude nous disposons à la fois d'une base de sons d'arabe standard, d'une base de durée de sons moyens pour tous les phonèmes, et de trois paramètres CS (coefficient de réduction syllabique), CR (coefficient de réduction phonémique) et CA (coefficient d'allongement phonémique).

Des études complémentaires sur la prise en compte de la structure accentuelle sont en cours afin d'établir une base de mesures prosodiques ainsi qu'une modélisation de contours prosodiques de l'arabe standard.

BIBLIOGRAPHIE

[Amr98] Amrouche A., Boudraa B., Rouvaen J.M. (1998), "Organisation temporelle des voyelles dans les structures CVCVCV, CVCCVCV et CVCCV de l'Arabe standard", JEP'98, pp. 91-94.

[Bar87] Barkova K., Sorin C. (1987), "A model of segmental duration for speech synthesis in french", Speech Communication, pp. 245-260.

[Bla76] Blachère R. (1976), "Eléments de l'arabe classique", Maisonneuve & Larose, Paris.

[Boe96] Boeffard O., Bigorgne D., Cherbonnel B., Emerard F., Roussarie L., Bagshaw P., Conkie A., Ennilo M., Traber C. (1996), "Utilisation des techniques d'apprentissage automatique pour les traitements linguistiques et prosodiques en synthèse de la parole : quelques résultats en Anglais, Allemand et Français", Actes des XXIèmes JEP, Avignon, France, 10-14 Juin, pp. 383-386.

[Dic98] Di Cristo A., Di Cristo P., Véronis J. (1998), "Optimisation d'un modèle prosodique pour la synthèse par règles à partir du texte du Français", JEP'98, pp. 135-138.

[Dut97] Dutoit T. (1997), "High-Quality Text-to-Speech Synthesis", Kluwer Academic Publishers.

[Gha92] Ghazali S., Znagui M., Benmiled Z., Jemni H. (1992), "Synthèse de l'arabe standard à partir du texte par TD_PSOLA : Le traitement des processus phonologiques", JEP'92, pp 89-93.

[Jom94] Jomaa M. (1994), "L'opposition de durée vocalique en arabe : Essai de typologie", JEP'94, pp. 395-400.

[Lec75] Lecomte G. (1975), "La Grammaire de l'arabe", P.U.F, Paris.

[Mal98] Malfrère F., Dutoit T., Mertens P. (1998), "Un générateur de prosodie 'Tout Automatique'", JEP'98, pp. 147-150.

[Nag98] Nagshaw P. C. (1998), "Unsupervised training of phone duration and energy models for text-to-speech synthesis", Proceeding of the 5th ICSLP, Sydney, Australie, 30 November-4 December 1998, pp 17-20.

[Mer99] Mertens P. (1999), "Un algorithme pour la génération de l'intonation dans la parole de synthèse", TALN'99, Cargèse, pp 233-242.

[Tou97] De Tournemire S. (1997), "Identification and automatic generation of prosodic contours for text-to-speech sysntem in french", Proceeding of EUROSPEECH'97, Rhodes, Grèce, 22-25 september 1997, Vol. 1, pp 191-194.

[Ros88] Rossi M. (1988), "Prosodies et technologies vocales ", Actes du GRECO-PRC, pp. 62-80.

[Zak84] Zakaria A.A, (1984), "L'accent Arabe Soudanais", Thèse de 3ème Cycle, Université de Franche-Conté, Besançon.

[Zem98] Zemirli Z. (1998), "SYNTHAR+ : Synthèse vocale arabe sous Multivox", TSI Vol17, N°6/98, pp. 741-761.

Rôle de la prosodie dans la communication en milieu bruité

Marie Dohalská, Jana Mejvaldová

Institut de Phonétique, Université Charles,
nám. J. Palacha 2, 116 38 Prague 1, République Tchèque
tél.: ++4202 216 19 250 – fax: ++420 24 81 21 66
marie.dohalska@ff.cuni.cz - mejvaldo@ff.cuni.cz
http://www.ff.cuni.cz/departments/fu/ip-intro.html

ABSTRACT

Our previous research in prosody of communication under degraded acoustic conditions showed that the segmentation of rhythm and melody, especially the placement and length of pauses as well as the characteristic contour, exercise an immediate influence on workers reaction and it may become virtually the only clue to draw attention to an emergency. We will extend the findings of our previous research in manipulation of these prosodic features on synthetic speech, particularly as related to specificities of "communication styles" .

1. INTRODUCTION

Dans les recherches sur la perception dans les conditions défavorisées, non standard (certains domaines professionnels – gestion dans les milieux bruités), il est souhaitable de se concentrer non seulement sur les phénomènes qui influencent la compréhension de l'information communiquée, mais également et surtout sur les facteurs qui contribuent à la fiabilité de la perception ou, au contraire, qui la perturbent.

Dans la communication quotidienne, la prosodie joue un rôle irremplaçable pour la compréhension d'un message sonore. La réalisation différente du groupement rythmique, de l'accent, de la courbe mélodique et d'autres phénomènes évoquerait un sens différent du message. Ainsi, la réalisation exacte de la prosodie dans les différents domaines de l'industrie, des mines, de l'agriculture ou du trafic (chemins de fer, trafic urbain, trafic aérien, etc.) est d'une importance particulière, parce que, comme une série de tests l'a prouvé, le contenu verbal n'est compréhensible, même pour les auditeurs spécialistes dans le domaine, qu'à 44 %. En déchiffrant les messages, ils s'appuient sur la connaissance du milieu et sur les mécanismes de compensation. Le répertoire des informations transmises est restreint, ce qui leur permet de s'orienter selon certains clichés sonores. L'information prosodique se montre plus robuste que l'information verbale. [DohZi88] La langue de la communication réelle doit assumer, malgré un haut degré de la déformation, la transmission relativement fiable des information nécessaires pour le fonctionnement du trafic.

Nos études de la communication dans les milieux plus ou moins bruités ont montré à quel point la réalisation concrète du rythme et de l'intonation, la distribution et la longueur des pauses, le changement inattendu du débit ou encore une accentuation inhabituelle orientent l'attention des communicants vers une situation urgente. Les changements de ce type sortent des stéréotypes prosodiques utilisés dans le milieu concret et contribuent d'une façon remarquable à la compréhension rapide d'un message urgent ou d'une question, mais ils peuvent également, comme nous avons eu la possibilité d'observer, avertir les employés d'une situation sérieuse non prévue et dangereuse.

2. MATERIAUX SONORES EXAMINES

Dans cette contribution, nous voulons présenter d'une part quelques caractéristiques typiques pour la prosodie utilisée dans la communication des postes de commande et d'autre part effectuer, sur ces matériaux, l'analyse de la réalisation prosodique de l'annonce d'un accident qui s'est produit à la gare de triage pendant le service nocturne.

2.1 Acquisition du matériel

Tous les messages décrits ci-dessous ont été enregistrés d'une façon professionnelle au cours de nos recherches concernants les différents types d'émissions des aiguilleurs en chef. L'enregistrement a été réalisé à l'aide d'un magnétophone professionnel, parallèlement sur les deux pistes de la bande magnétique professionnelle classique. La première piste enregistrait le message sonore en direct, tel que les cheminots l'avaient entendu, et la piste parallèle enregistait le signal pris à la source du microphone du chef du poste de nuit. La durée totale des matériaux sonores transcrits était 240 minutes. Etant donné qu'il s'agit strictement du matériel réel, les locuteurs sont les employés du trafic concret.

L'enregistrement parallèle était indispensable pour la compréhension du contenu verbal à cause des mauvaises conditions acoustiques et de l'articulation très négligée des locuteurs. L'intelligibilité des émissions en direct n'atteignait que 4,8 % pour les non-spécialistes; la compréhension des employés a été liée à la situation concrète; après quelque jours et ne disposant plus du

contexte réel, ils n'ont pas été capables de répéter l'entendu.

3. CLICHES DE COMMUNICATION

Dans le domaine de travail observé, les messages fréquents se sont plus ou moins stabilisés sous formes de «clichés de communication». Dans le cas de la gare de triage (nous avons eu la possibilité de suivre et d'analyser la communication dans trois gares différentes), nous avons pu retenir deux types principaux:

a) phrases contenant presque uniquement des chiffres („phrases-chiffres");

b) „phrases-instructions" qui décrivent le travail à effectuer pour un groupe entier d'employés.

Phrases-chiffres Les phrases où dominent les chiffres contiennent, en moyenne, de 14 à 16 syllabes et elles sont divisées d'habitude en trois groupes respiratoires dont la longueur est donnée par la composition de chiffres. L'organisation temporelle est régulière, l'intonation pratiquement monotone descendante à la fin. La durée des phrases de ce type varie en général dans un intervalle de 4 à 5 secondes.

Phrases-instructions Les groupes respiratoires dans ce type de phrases contiennent en moyenne de 24 à 26 syllabes. Leur organisation temporelle et répartition syllabique sont moins régulières, à la fin des groupes on trouve des expressions de prise de contact ou les noms des auditeurs juxtaposés. La durée et la structure mélodique et rythmique des phrases-instructions varient très peu et très rarement.

Type de discours informatif	Durée totale	Nombre et No de groupes	Durée (ms)	Nombre de syllabes	Débit (syll/s)	Durée des pauses (ms)
	9509 ms	1.	662	3	4,5	128
		2.	3946	24	6,2	1115
		3	3896	25	6,4	-

Avis de danger	Durée totale	Nombre et No de groupes	Durée (ms)	Nombre de syllabes	Débit (syll/s)	Durée des pauses (ms)
	9509 ms	1.	2604	7	2,7	317
		2.	701	2	2,9	243
		3.	1524	5	3,3	451
		4.	5328	33	6,2	697
		5.	1800	10	5,6	566
		6.	2751	12	4,4	-

Tab. 1: Caractéristiques rythmiques de deux types de phrase

4. PHRASES CHIFFRES

Phrases chiffres sont composées presque uniquement des chiffres indiquant l'endroit et le type de travail à faire. Le message sonore réel, tel qu'il a été enregistré pendant notre expérience, a été constitué par ce que nous appelons "patrons vocaliques" et qui n'est qu'une succession des voyelles sans aucune possibilité de déchiffrer le contexte consonantique.

Ex.: na dvanáctou / ve dvacet jedna se transforme en
a a a: ou/ e ae e a

Les consonnes ne sont pas distinguées; c'est leur regroupement et surtout les tenues des occlusives qui aident à reconstruire le message émis.

4.1 «Patrons vocaliques»

Les messages émis sont donc réduits à des simples successions de voyelles, «patrons vocaliques». Cette stratégie de perception peut conduire, bien évidemment, à des confusions qui empêcheraient l'accomplissement de la tâche demandée. Seule la connaissance du contexte situationnel permet la compréhension du message et la réaction par un travail adéquat. [Ba96] Quatre interprétations différentes ont été trouvées dans les tests perceptifs; chaque variante était admissible pour le type d'activité, mais une seulement correspondait au travail demandé. Malgré la stabilité relative des voyelles, elles ont subis quelques types de confusion; le tableau suivant montre les directions principales de la perception vocalique:

Les voyelles [i], [a] et [u] se montrent plus stables que [e] et [o].

voyelle	peut être remplacée par	exceptionnellement par
[i]	*[e]*	*[u]* ou *[a]*
[e]	*[a]* ou *[o]*	*[i]*
[a]	*[e]* ou *[o]*	
[o]	*[a]*, *[e]* ou *[u]*	
[u]	*[o]*	*[i]* ou *[e]*

Tab. 2: Types de confusion vocalique.

4.2 Changements consonantiques

Les occlusives [p], [m], [g] peuvent être remplacées par une autre occlusive. La variation perceptive plus grande existe pour les consonnes (surtout dans la position initiale) [t], [d], [n]: elle peuvent être remplacées par d'autres occlusives mais également par les constrictive [s], [z] et [j]. La dernière a souvent remplacé les occlusives palatales. [x] a été remplacé par les occlusives [t], [k], [n] uniquement, [s] a été perçu comme [ʃ], [x], [z], exceptionnellement comme [t]. Quant aux consonnes afriquées, [ts] s'est montré plus stable que [tʃ] qui a été remplacé par [ʃ], [s], [j], [m] et autres.

Dans les exemples présentés ci-dessus, nous avons montré quels sont les types de substitution rencontrés; mais l'objectif de notre travail n'est pas de présenter un aperçu des changements perceptifs: il est nécessaire de prendre en considération la structure de chaque mot, son découpage syllabique, la fonction et la position de l'accent etc. Dans la communication verbale quotidienne l'auditeur complète automatiquement, dans la plupart des cas, les informations apportées par la chaîne sonore grâce à la connaissance de la situation communicative.

5. PHRASES –INSTRUCTIONS

Pour présenter le type de la phrase-instruction, nous avons choisi l'exemple suivant (83-1.wav) :

A Jindro, (une apostrophe)

[a'jindro] 662 ms 4,5 sl/s

potom az pojedeš tak tady u toho blafounu pomalu

['potomaʃ' pojedeʃtak 'tadi 'utoho 'blafounu 'pomalu

abych se moh chytnout. 3946 ms 6,2 sl/s

['abixsemox 'xitnout]

(demande au conducteur de ralentir devant le haut-parleur pour que l'autre cheminot puisse monter en marche; la traduction exacte n'est pas possible, car il s'agit d'un jargon professionnel très peu compréhensible même dans la langue maternelle aux auditeurs profanes).

L'instruction suivante explique qu'il faut s'arrêter là où le wagon sera déchargé dans la nuit:

Protože já bych mu tam zastavil jak von chce

['protoʒe' ja:bixmutam'zastavil' jakvonxce]

von tam bude vykládat asi v noci víš

['vontambude 'vikla:datasivnoci'vi:ʃ] 3896 ms 6,4 sl/s

5.1 Caractéristiques rythmiques et temporelles de la phrase-instruction

Le premier enregistrement est un exemple du discours informatif: pratiquement tous les messages de ce type ont été construits, sans que le locuteur le fasse consciemment, de la même façon: la même longueur du message (la possibilité de se concentrer sur les messages plus étendus étant très limitée), la même distribution des pauses, la même longueur des groupes à l'intérieur du message (dans cet article, nous appelons «groupe» la partie du discours délimitée par les pauses). La longueur des groupes ne varie pas – le premier groupe, de trois syllabes, est une apostrophe et ici seulement le débit est différent. [LeRo80]

Vu la qualité du signal enregistré dans de mauvaises conditions acoustiques (gare de triage), il n'était pas possible d'effectuer une analyse expérimentale complète de la fréquence fondamentale et de l'intensité. Néanmois, l'analyse auditive appuyée sur les résultats d'une série des tests auditifs a permis de constater la monotonie mélodique (absence de variations entre les syllabes successives), avec la tendance remarquablement continue vers la descente, sans aucune déviation. Par rapport à la prononciation tchèque standard, les groupes rythmiques sont plus longs – plusieurs groupes sont reliés en un groupe accentuel.

6. PROSODIE «DE L'ACCIDENT»

Le cliché prosodique que nous venons de décrire est essentiellement différent de la prosodie réalisée dans le but d'annoncer une situation dangereuse, où plusieurs wagons se sont détachés et descendaient, sans pouvoir être contrôlés rapidement, la pente de la gare (la structure rythmique est décrite et comparée dans le tableau 1). L'aiguilleur en chef devait absolument avertir les cheminots du danger sérieux imprévu. La qualité du signal était très basse et le contenu verbal incompréhensible pour les cheminots se trouvant sur les voies. Dans cette situation, seule la prosodie alarmante a assumé la transmission de l'information nécessaire.[Doh-Zi91]. Le texte décrivant la situation grave de cet accident inattendu est le suivant (83-2.wav):

Heehee pozor tam dole

['he:e:' pozortamdole] 2604 ms 2,7 sl/s

Honzo

['ho:nzo:] 701 ms 2,9 sl/s

Honzo do pree

['honzo:' dopre:e:] 1524 ms 3,3 sl/s

Honzo pozor vodpojilo se to, jak to cuklo odpojilo se to

['honzo'pozor 'votpojiloseto' jaktocuklo' odpojiloseto]

a bezí asi sedum nebo osum vozu

['abjeʒi:asi 'sedum'neboosum'vozu:] 5328 ms 6,2 sl/s

ale vzadu to má kliku vole

['alevzdutoma:'klikuvole] 1800 ms 5,6 sl/s

ten poslední štyrosák má kliku boudu.

*[ten'posledni:'ʃtirosa:kma: 'klikuboudu]*2751 ms 4,4 sl/s

Ce message peut être divisé en deux parties selon la fonction qu'elles assument: dans la première partie, le locuteur doit attirer l'attention des cheminots (dont un est en train de parler en continuant son message, appartenant au type prosodique des phrases-instructions).

La première partie (séparée, dans le tableau 1, par la ligne en grasse) est alors construite de mots de prise de contact (apostrophes, etc.). Elle est caractérisée par l'intensité extrêmement forte et surtout par la mélodie montante en deux pics atteignant les valeurs de 290 Hz et 297 Hz.

La deuxième partie est moins agitée. La mélodie dans son dernier groupe rythmique descend de 240 Hz à 157 Hz. L'intensité reste très forte dans toute la phrase et suit, à la fin, la courbe de la fréquence fondamentale.

L'information a été transmise uniquement grâce à l'organisation inhabituelle du message: par rapport aux phrases-instructions, la longueur des groupes respiratoires est irrégulière, le débit variable et l'information verbale (presque incompréhensible) est concentrée dans un groupe d'une longueur extrême. Les deux parties de ce message assument deux fonctions et leurs structures sont, en accord avec ce fait, différentes. Par exemple, la proportion des pauses dans la première partie est de 1011 ms par rapport aux 4829 ms de phonation, ce qui représente 17%; dans la deuxième partie cette relation est de 1263 ms: 9879 ms (11 %).

7. CONCLUSION

Nos expériences basées sur les analyses de la parole professionnelle de certaines branches de l'industrie ou du trafic (gares de triage, chemins de fer, trafic urbain, etc.) montrent que la communication professionnelle dans le milieu bruité est représentée dans la plupart des cas par des clichés d'une structure rythmique et mélodique stable. Le contenu verbal des phrases a, en général, une valeur informative très faible causée non seulement par le bruit, mais aussi par la prononciation négligée des locuteurs. Ce type de communication professionnelle n'aurait pas pu, dans la plupart des cas, être adopté sans l'existence de mécanismes de compensation qui permettent aux employés de déchiffrer le sens des instructions ou des questions d'après leurs expériences professionnelles [Du81]. Dans le cas de l'accident décrit ci-dessus, le texte était entièrement indéchiffrable. Nous le répétons, seules la répartition rythmique atypique, les pauses fonctionnelles et importantes, ensemble avec la mélodie très montante au début de la phrase, qui, par sa forme prosodique, sont tout a fait différentes des clichés informatifs courants dans le tchèque standard [Palk94] mais aussi dans le type concret du trafic, ont apporté le maximum d'information sur une situation très dangereuse [Fo91], pratiquement en absence de l'information verbale.

La structure de la répartition syllabique, de l'accentuation (accent tonique et dynamique), de l'intonation et du débit caractéristique des clichés informatifs standards est très différente de la structure des informations impérieuses, urgentes. C'est pourquoi nous voulons essayer de formaliser des types des mises en garde et des avertissement dans le tchèque synthétique [DohMej97]. La forme synthétique, liée aux détecteurs automatiques contrôlant plusieurs types de danger, serait très efficace, car les traits prosodiques caractéristiques pour les situations dangereuses sont liés, dans la parole naturelle, à l'articulation déformée des segments. Les avertissements synthétiques modélisés selon la prosodie réelle seraient non moins suggestifs par leur forme prosodique, mais en plus ils seraient plus claires et efficaces par la régularité segmentale de la parole synthétique. Inspirées par le projet COST 258 et par l'étude des «styles» qu'il englobe, nous pensons à enrichir les styles déjà définis par ce type «d'avertissement».

BIBLIOGRAPHIE

[LeRo80] Léon, P., Rossi, M. (1980) Problèmes de Prosodie, Studia Phonetica 17, Didier Ottawa

[Du81] Durand, J. (1981) Les formes de la communication, Bordas, Paris

[Doh-Zi88] Dohalská-Zichová, M. (1988) Contribution à l'étude de la perception de la parole dans de mauvaises conditions acoustiques, Phonetica Pragensia VII, UK, Prague

[DohZi91] Dohalská-Zichová, M. (1991) Dynamika verbální komunikace, AUC UK Praha

[Fo91] Fónagy, I. (1991) La vive voix, Essais de psycho-phonétique, Payot, Paris

[Palk94] Palková, Z. (1994) Fonetika a fonologie češtiny, Karolinum, Praha

[Ba96] Baylon, Ch. (1996) Sociolinguistique, Société, Langue et Discours, Nathan

[DohMe97] Dohalská-Zichová, M., Mejvaldová, J. (1997) Où sont les limites phonostylistiques du tchèque synthétique, XVI^{e} CIL, Paris

Les voyelles toniques des paroxytons francoprovençaux

[1]S. Roullet & [2]L. Molinu
[1]Centre de Dialectologie de l'Université Stendhal
Grenoble III – BP 25 – 38040 Grenoble, France
[2]Université Toulouse Le Mirail
5, Allées A. Machado – F –31058 Toulouse Cedex 1
Tél.: ++33(O)561 50 36 93 Fax: ++33(O)561 50 46 77

Mél: molinu@univ-tlse2.fr

ABSTRACT

Notre article se propose d'analyser la réalisation des voyelles toniques dans des mots paroxytoniques d'une variété francoprovençale. Nous avons cherché à expliquer la distribution de la durée vocalique et consonantique dans les syllabes. toniques en utilisant les traitements proposés par les théories phonologiques les plus récentes.

1 INTRODUCTION

Notre analyse se propose d'étudier la réalisation des voyelles toniques dans des mots paroxytoniques, prononcés dans la variété francoprovençale parlée à Sarre, que nous estimons représentative du "valdôtain central" [Rom98].
Elle se base tout particulièrement sur des mesures de durée menées sur 134 mots. Pour chacun d'entre eux nous avons pris en considération 3 répétitions, de manière à réduire, même si partiellement, l'incidence des variations dues à des changement intervenus dans le débit. Par la suite, sur la base des données obtenues, nous avons cherché à expliquer quelques phénomènes particulièrement intéressants, concernant la distribution de la durée vocalique et consonantique dans les syllabes toniques. Pour ce faire, nous avons considéré "l'histoire étymologique" des mots et nous avons utilisé des traitements proposés par les théories phonologiques les plus récentes [Par88].

1.1 Phénomènes relevés dans la variété francoprovençale parlée à Sarre

Il a été démontré que plusieurs phénomènes de réduction et de renforcement se trouvent souvent en relation avec l'absence ou la présence de l'accent : les syllabes non accentuées manifestent la tendance à engendrer des réductions, tandis que les syllabes accentuées sont à l'origine de renforcements [Str64 ; Nes93]. On relève, en outre, que le contraste entre syllabe ouverte et syllabe fermée en position tonique peut modifier la quantité et la qualité de la voyelle associée au noyau syllabique. Ainsi, dans beaucoup de langues, on a mis en évidence un allongement des voyelles accentuées seulement en syllabe ouverte.
Un phénomène analogue a été relevé également dans la variété francoprovençale que nous avons prise en considération ; sur la base des mesurations que nous avons effectuées, on remarque que la voyelle accentuée, en syllabe fermée, est toujours brève, tandis qu'elle est longue si la syllabe est ouverte :

['sa:la]	(salle)
['pap:a]	(papà)
[rə'ko:sə]	(regain)
[tsa'mɔs:ə]	(chamois)

La variété que nous avons analysée se caractérise également par une distinction qualitative dans la réalisation des voyelles moyennes [e] et [o], en syllabe fermée et en syllabe ouverte. Dans le premier cas, en effet, la voyelle présente toujours un timbre plus ouvert par rapport à celui qui caractérise autrement la voyelle. Cette variation qualitative n'est pas relevée en syllabe atone ; dans ce cas, le noyau syllabique est constitué par une voyelle moyenne fermée.

[tsa'rɛt:a]	(charrette)	[tsare'ti]	(charretier)
['dɔb:lo]	(double)	[do'blu]	(redoublé)

En syllabe ouverte, nous n'avons pas relevé des changements importants du timbre vocaliques, engendrés par le déplacement de l'accent.
Par conséquent, nous pouvons affirmer que la structure métrico-syllabique exerce un'influence considérable sur la réalisation du timbre des voyelles : l'ouverture des voyelles moyennes est conditionnée par la position de l'accent et par la réalisation d'un consonne associée en position de coda [Mar94].

2 CONSIDERATIONS D'ORDRE HISTORIQUE ET ETYMOLOGIQUE

Une analyse plus approfondie, dédiée tout particulièrement à l'étude de l'évolution des mots au cours des siècles dans la variété francoprovençale considérée, nous a permis de montrer que, si la tonique est issue d'une longue latine, le trait [+ long] ne concerne pas la voyelle, mais la consonne suivante. Dans ce cas, en effet, la "force" qui caractérisait la voyelle latine dans les aboutissants ne s'est pas "concentrée" sur la seule voyelle tonique, mais elle s'est portée sur la consonne suivante, en l'allongeant :

R O M A

['r ɔ m : a]

Le phénomène concerne exclusivement les syllabes toniques ; lorsque l'accent se réalise sur la syllabe suivante, la gémination ne se produit pas.
En revanche, si la voyelle de la syllabe tonique dérive d'une voyelle brève latine, elle présente le trait [+ long] :

R O S A

↓

['r u : z a]

2.1 *Les dissyllabes*

En ce qui concerne les dissyllabes, nous pouvons schématiser leur évolution de la manière suivante :
1 - si la voyelle tonique latine se trouvait en *syllabe fermée*, la voyelle francoprovençale est **brève** :

PORTA ['purta]

2 - si la voyelle latine était *longue en syllabe ouverte*, elle a abouti à une **brève**, suivie par une consonne "longue" :

RIPA ['riv:a]

3 - les *diphtongues* toniques latines ont également abouti à une voyelle **brève** (la consonne suivante a été "longue"):

PAUCUS ['pɔk:a]

4 - si la voyelle latine était une *brève en syllabe libre*, elle est devenue une voyelle **longue** (par conséquent, la consonne suivante est brève) :

POLUS ['po:lə]

2.2 *Les paroxytons formés par plus de deux syllabes*

La majorité des paroxytons formés par plus de deux syllabes présente une structure de type VCCV. Elle caractérise évidemment les mots francoprovençaux qui dérivent de paroxytons latins, dont la voyelles tonique était en syllabe fermée ou s'il s'agissait d'une longue en syllabe ouverte, mais aussi des proparoxytons, dont la pénultième syllabe était sûrement brève et, par conséquent, atone :

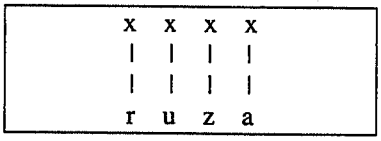
FABRICA [fa'brək:a]
MACHINA [ma'ʃin:a]

Les données dont nous disposons laissent supposer qu'à un moment donné même les mots latins proparoxytoniques ont été prononcés comme des paroxytons ; par conséquent, ils auraient subi les mêmes évolutions. Une confirmation de notre hypothèse nous la trouvons déjà chez P. Gardette [Gar83], qui faisait remarquer que « *le francoprovençal conservait plus fidèlement la forme du mot latin en transportant l'accent sur la pénultième, du moins dans un certain nombre de mots* ».

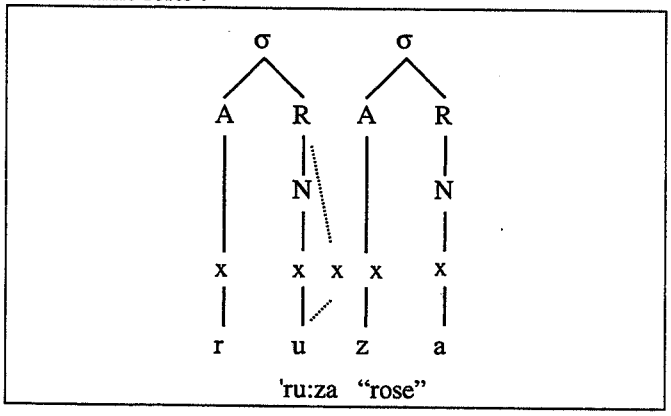
3 ANALYSE PHONOLOGIQUE

Nous avons concentré notre analyse sur les alternances consonne géminée *vs.* consonne simple (C:/C) et voyelle longue *vs.* voyelle brève (V:/V). Comme nous l'avons montré dans les paragraphes précédents, ces alternances sont conditionnées par la structure métrico-syllabique.

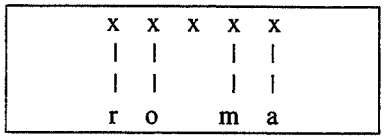
Pour chercher à expliquer ce phénomène, nous nous sommes basées sur des considérations d'ordre diachronique et nous avons utilisé des instruments théoriques, qui ont été proposés par les plus récentes théories phonologiques. Plus particulièrement, nous nous sommes inspirées de la *Théorie des contraintes et des stratégies de réparation* ; il s'agit notamment d'un modèle phonologique plurilinéaire qui interprète les différents processus phonologiques comme étant le résultat de 2 opérations fondamentales (les stratégies de réparation) : l'insertion et l'élision [Par88]. Ces 2 opérations s'expliquent par l'interaction, dans la grammaire de toute langue, de principes généraux et de restrictions paramétriques.
La gémination consonantique et l'allongement vocalique en syllabe tonique sont deux processus qui ont en commun la nécessité de respecter la restriction sur la rime forte [Chi86 ; Mar95]. Leur application **C :** *vs.* **V :** est déterminée par la structure lexicale des "mots" qui doivent être accentués ; à notre avis, elle est restée fidèle, du moins à certains niveaux de la représentation phonologique, à la forme étymologique.
Dans le cas où se réalise l'allongement vocalique, sur la base de considérations d'ordre étymologique, nous supposons une représentation sous-jacente de ce type:



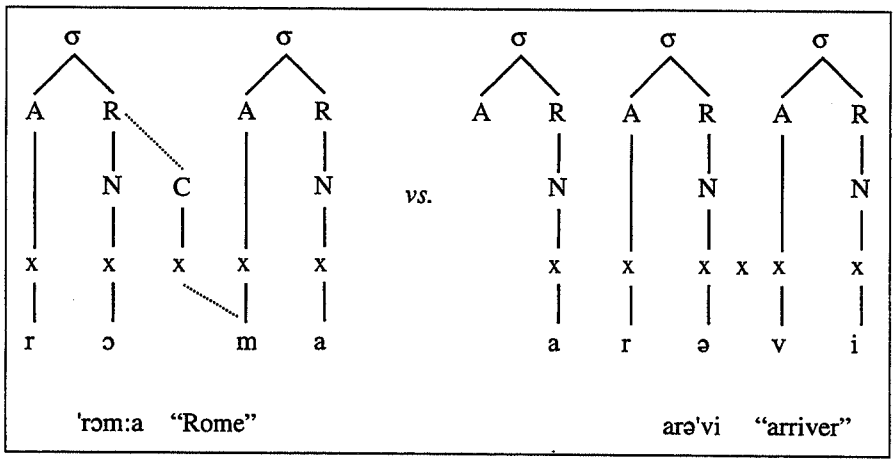
La forme sous-jacente n'a pas subi de modifications par rapport à la forme originaire : le nombre des unités temporaires et des segments n'a pas changé. Les modifications interviennent dans la dérivation et elles sont engendrées par les conditions métrico-syllabiques : si la voyelle est accentuée, nous assistons à l'insertion d'une unité chronématique et d'une ligne d'association qui unit le *slot* à la voyelle. Cette opération engendre la réalisation de la position temporelle par l'allongement de la voyelle et, par conséquent, elle permet de respecter la restriction sur la rime forte :



En revanche, dans le cas où se réalise la gémination consonantique, nous proposons la structure lexicale suivante :



Dans ce cas, il y a eu une réinterprétation de la forme étymologique qui était caractérisée par une voyelle longue (RO:MA) : l'unité chronématique a été dissociée, au cours de l'évolution, de l'élément vocalique et elle a acquis le statut de élément fluctuant. La modification a donc intéressé le niveau segmental, mais pas celui de la squelette constituée par des unités chronématiques. L'ancrage du *slot* à la consonne, permis par l'insertion d'une ligne de dissociation, est déterminé encore une fois par l'accent : si la syllabe est accentuée l'association se vérifie ; dans le cas contraire l'élément reste fluctuant et il ne reçoit aucune interprétation phonétique :



L'utilisation d'un modèle comme la **TCSR** nous a permis non seulement d'interpréter la gémination consonantique et l'allongement vocalique comme étant le produit de l'insertion de matériel phonologique, mais aussi d'expliquer cette opération sur la base de la restriction sur la rime forte.

Malheureusement, notre analyse n'est pas tout à fait satisfaisante. Nous n'avons pas pu rendre compte des conditions qui déterminent le choix entre la gémination consonantique et l'allongement vocalique. Pourquoi y a-t-il gémination quand la forme dérive d'une structure de type CV:CV, e pourquoi l'allongement vocalique se produit-il dans des syllabes tonique qui en latin présentaient une voyelle brève ?

Nous estimons donc que d'autres recherches seraient nécessaires, pour qu'elles puissent permettre de trouver un réponse à ces problèmes.

BIBLIOGRAPHIE

[Chi86] Chierchia G., Length, *Syllabification and the Phonological cycle in Italian*, Journal of Linguistics, 8, 5-33, 1986.

[Gar83] Gardette P., Études de géographie linguistique, ed. par B. Horiot, M.R. Simoni & G. Straka. Klincksieck, Strasbourg, 1983.

[Mar94] Marotta G. & Savoia L. (1994). *Vowel properties and nuclear constituents: Evidence from Italian dialects*, Probus, 6, 43-79, 1994.

[Mar95] Marotta G., *La sibilante preconsonantica in italiano: questioni teoriche e analisi sperimentale*, Scritti linguistici e filologiciin onore di Tristano Bolelli, Pisa, Pacini, 393-438, 1995.

[Nes93] Nespor M., Fonologia. Bologna, Il Mulino, 1993.

[Par88] Paradis C., *On Constraints and Repair Strategies*, The Linguistic Review, 6, 71-97, 1988.

[Rom98] Romano A. & Roullet S., *Analisi intonativa comparata di due varietà di italiano regionale (salentino meridionale e valdostano centrale) sulla base di un corpus fisso di frasi affermative e interrogative, ottenute mediante progressive espansioni dei sintagmi nominale e verbale*, in Unità fonetiche e fonologiche: produzione e percezione. Atti delle VIII Giornate del Gruppo di Fonetica Sperimentale – A.I.A. (Pisa, 18-19 dic. 1997), ed. par P.M. Bertinetto & L. Cioni. Pisa, Scuola Normale Superiore, 128-141, 1998.

[Str64] Straka G., *L'évolution phonétique du latin au français sous l'effet de l'énergie et de la faiblesse articulatoire*, Travaux de Linguistique et de Littérature, II, 1, Univ. de Strasbourg, 17-98, 1964.

Les tons comme voie d'accès au lexique : le cas des dérivés initiatiques ohendo

Hubert NGONGA-ke-MBEMBE

Laboratoire de Phonologie – Université Libre de Bruxelles
Tél.: ++32 (2)650 20 18 - Fax: ++32 (2)650 20 07
http://www.ulb.ac.be/philo/phonolab

ABSTRACT

This paper shows that tone can play an important role in lexical identification in Ohendo. Evidence come from word games performed in initiatory languages performed during initiation. We conclude that tone, as other pre-lexical units, are intermediate representations for spoken word recognition.

1. INTRODUCTION

Cet article examine la manière dont les tons peuvent constituer une voie d'accès au lexique dans les mots dérivés ohendo. Le lohĕhendo est une langue bantoue parlée en République Démocratique du Congo dans la Région du Kasaï-Oriental, plus précisément dans la zone de Kole. Le système phonologique ohendo comprend 16 consonnes simples [p,b,t,d,k,m,n,ɲ,ŋ,f,s,ʃ,ʒ,h,j,l], 21 consonnes complexes et 7 voyelles [i,e,ɛ,a,u,o,ɔ]. Le lohĕhendo a deux tons simples : un ton bas et un ton haut. La combinaison de ces deux tons donne lieu à des tons modulés montant et descendant.

2. DONNEES

La formation de langues initiatiques recourt à plusieurs procédés dont la dérivation qui acquiert un sens tout à fait spécial dans ce contexte: on peut la définir comme une altération systématique de la forme de la langue courante, qui recourt à des mécanismes comme la troncation, la mutation, l'adjonction, la substitution... Les dérivés initiatiques sont des mots qui sont affectés par une, ou conjointement par deux ou plusieurs opérations qui forment le système sur lequel sont constituées les langues parlées exclusivement dans des associations initiatiques. Une langue de ce type est appelée dérivée initiatique (LADI). Sa forme varie en fonction des opérations dérivationnelles qui y sont appliquées. La table 1 montre des exemples d'opérations dérivationnelles dans trois mots ohendo. Les sigles L1-L19 renvoient aux langues initiatiques qui sont connues sous le terme générique de *lɔnɨnɨmi* 'langue puissante'. Comment les initiés, qui sont tous illettrés, procèdent-ils pour accéder au lexique (toutes les LADI sont orales)? Comment font les auditeurs initiés pour traiter les indices correspondant aux représentations lexicales abstraites dans le continuum acoustique? Deux problèmes importants sont à résoudre pour répondre à ces questions.

Le premier problème est que même s'il existe des indices phonétiques, phonologiques, phonotactiques et prosodiques qui indiquent des frontières potentielles, ceux-ci ne suffisent pas pour fournir une analyse lexicale assez claire et non ambiguë. A cet effet, les résultats de l'étude de Tabossi [Tab93] sont assez significatifs: ils montrent que les mots enchâssés peuvent être activés malgré la présence des indices de segmentation clairs.

Table 1. Exemples d'opérations dérivationnelles dans les LADI ohendo

ohendo	losálá	dʒăsa
Français	*une plume*	*un jumeau*
LADI: 1	lísofísalísa	dʒísafísa
2	lonɔ́sanálaná	dʒanásaná
3	losósa:láta	dʒasásáfa
4	lalásá	sása
5	losaláta	dʒasáfa
6	loŋgósaŋgálaŋgá	dʒaŋgásaŋgá
7	losalédʒa	dʒaśédʒa
8	lasáló	săʤa
9	lasólá	dʒĕsa
10	lolásá	săʤa
11	losátá	dʒăfa
12	losɛ́lɛ́	dʒĕsɛ
13	lósa:lă	dʒásâ
14	lokílá	dʒáki
15	masálálo	măsadʒi
16	loláfá	făʤa
17	lovaláta	dʒăvala
18	sa: lo	sa: i
19	sa:salo	sa:sai

Le second problème est que les LADI perturbent certains éléments servant de repères pour la reconnaissance de mots parlés, comme c'est le cas du *point d'unicité* proposé par le modèle *cohort*. L'hypothèse selon laquelle le point de reconnaissance du mot correspondrait à son point d'unicité devient caduque si on s'en tient à la forme de départ. L'altération que subit le mot courant est telle que la forme qu'il acquiert n'est compatible avec aucune donnée lexicale, au point qu'on ne peut envisager ni point d'unicité ni point de déviation. On peut juste considérer ces dérivés comme des entrées lexicales formellement nouvelles.

Cette étude propose que l'accès au lexique ne suit pas toujours le modèle du mot écrit. Elle montre que ceci peut être accompli grâce au rythme et au symbole non verbal.

3. LES REPERES DE RECONNAISSANCE DES MOTS

Dans la littérature, on constate que l'information dont se servent les différents procédés de reconnaissance de mots provient de deux sources: interne et externe. La première est relative aux propriétés du mot: la longueur [Meh68], [Gro80], la fréquence [Rub68], le point d'unicité et l'aspect phonotactique. La seconde est afférente au contexte et aux connaissances connexes. Ces procédés bifaces sont insensibles à la différence oral/écrit alors que les deux modes ont chacun des particularités dont il convient de tenir compte. Les LADI sont orales, on a donc affaire à un discours continu caractérisé par un signal de la parole qui se déroule dans le temps et par une absence de frontières clairement établies entre les unités linguistiques.

On sait qu'il existe des indices phonétiques, prosodiques, phonotactiques et phonologiques qui indiquent les frontières potentielles voire même probables [Chu87], mais ils ne sont pas suffisants pour fournir une analyse lexicale claire et non ambiguë. Tabossi [Tab93] a montré que les mots enchâssés peuvent être activés malgré les indices de segmentation clairs. Malgré ce fait, de nombreux modèles de reconnaissance de mots parlés acceptent que l'accès au lexique soit fonction de l'extraction préalable des unités linguistiques pré-lexicales: trait, segment, syllabe... Deux raisons rendent cette hypothèse plausible, comme le montrent Moraïs et al. [Mor91].

Les mots étant des objets phonologiques complexes, l'extraction des composantes phonologiques à un ou à plusieurs niveaux d'information constitue la voie de reconnaissance des mots la plus adaptée à la structure de la langue. Cette dernière atteste plusieurs régularités phonologiques, ce qui permet d'exclure un grand nombre de possibilités combinatoires. Il serait donc anormal que ces régularités ne soient pas prises en compte dans le système de reconnaissance des mots.

Les variations dues au bruit dans le signal, au locuteur, et au débit de parole sont telles que l'on voit mal comment les patrons spectraux pourraient directement être associés à la représentation des mots. Une segmentation qui se ferait en se basant sur les formes pré-lexicales a plus d'avantages que celles qui se ferait sur les formes spectrales. D'une part parce qu'on aura un nombre très réduit d'unités et d'autre part parce qu'on aura un nombre infini de formes spectrales.

S'il est évident que l'objectif de la segmentation consiste à retrouver les unités lexicales abstraites, rien n'indique que le traitement du signal se fasse par des unités abstraites inférieures ou égales au mot et non à partir des portions de signal supérieures au mot ou définies par des critères autres que des critères orthographiques. C'est dans cet esprit qu'il convient d'examiner la segmentation des dérivés initiatiques.

4. LA SEGMENTATION DES DERIVES INITIATIQUES

L'examen des items repris dans la table 1 montre que les cibles visées par les opérations dérivationnelles révèlent une segmentation naïve des locuteurs. En même temps, on peut observer que les différents mécanismes altèrent la langue courante. Selon qu'ils opèrent en singleton ou conjointement, la distinction est faite entre une langue dérivée simple et une langue dérivée composée. Une même opération peut, au cours d'une dérivation donnée, cibler deux unités linguistiques de taille différente comme le segment et le trait. C'est le cas de l'adjonction qui apporte des additifs segmentaux grâce à la propagation ou en manipulant les traits en recourant à la dissimilation.

Cette extraction naïve mais consciente des unités linguistiques pré-lexicales pose deux problèmes.

Elle peut être morphologiquement localisée au début, au milieu ou à la fin d'un mot. Elle peut concerner deux ou toutes les positions à la fois, ainsi qu'on le remarque respectivement dans L5, L15 et L8. On peut dès lors se poser la question de savoir laquelle des positions reste déterminante pour l'identification de l'unité lexicale.

Elle peut aussi être faite sur deux unités linguistiques de niveaux différents, comme la syllabe et le segment, qui peuvent être morphologiquement situées à des positions différentes comme dans les L3, L17. On peut ici se poser la question de savoir laquelle de deux unités extraites est déterminante quant à l'accès au lexique et quel est son rapport avec la position qu'elle occupe dans le mot.

Ces données supposent que la reconnaissance des dérivés initiatiques implique une analyse acoustico-phonétique préalable du signal de la parole. Il est indispensable que les frontières établies par cette analyse introduisent une distinction entre relation déterminante et non déterminante, lorsqu'il faudra définir le rapport entre les différentes entités du signal et les représentations lexicales abstraites stockées dans le lexique mental. Le caractère déterminant d'une relation est exprimé par les propriétés saillantes de celle-ci. Cette relation couvre plusieurs phénomènes, les tons, la durée, la structure syllabique: lorsqu'il est en rapport avec les syllabes, le caractère saillant tend à porter ce qui a le plus de sens et qui est le plus redondant dans la phrase. Ce travail montre que la conjonction des tons, de l'accent et de symboles rend possible un accès lexical rapide.

5. LA PROSODIE COMME VOIE D'ACCES AU LEXIQUE

Le terme rythme en linguistique réfère autant à la théorie métrique qu'à la phonologie prosodique. Les deux modules prennent en compte la prosodie en tant qu'ensemble des éléments suprasegmentaux parmi lesquels les tons jouent le rôle essentiel. Le rapport asymétrique *fort/faible* de l'organisation hiérarchique est favorable à l'entrée lexicale. De nombreux auteurs soutiennent que dans les langues qui présentent une stratégie de segmentation métrique comme l'anglais et le néerlandais, les syllabes fortes facilitent l'accès au lexique

[Cut88]. Dans le cas ohendo, il s'agit des dérivés qui sont des formes altérées de la langue de base. Sur le plan strictement tonal, les paradigmes ohendo peuvent être regroupés selon des catégories tonales. Par ce terme, on entend le nombre de langues dérivées qui partagent un schème tonal souvent artificiellement créé dans le but d'occulter les tons réels en modifiant leur structure. Dans ce cas, ils gardent leur forme et leur position en changeant d'unités porteuses. On distingue ainsi la catégorie tonale imposée qui est caractérisée par un schème tonal indépendant et invariable. C'est dans ce cas que le rythme joue le rôle de la segmentation lexicale qui conduit à la reconnaissance de mots. Les tons sont toujours affectés de façon significative: (les tons hauts indiquent des éléments distrayants pour les langues additives) ce sont toujours des syllabes ou des infixes asémantiques. Les tons bas sont portés par des formes de la langue de base, ainsi qu'on l'observe dans L1, L2 et L6. Dans les trisyllabes où les dissyllabes, les tons attestés constituent le *mouvement*, c'est-à-dire une séquence de tons alternant le haut et le bas ou inversement. Ce peut être formulé de la manière suivante: [H B]n ; [B H]n : les tons haut et bas couvrent le mot dans une suite alternante en commençant soit par le ton haut soit par le ton bas.

Dans un contexte syntactique, les tons qui sont au début du mot suivant sont modifiés de la manière suivante:

a. losálá > loŋgósaŋgálaŋgá

B H B H B H

b. losálá lómi > loŋgósaŋgálaŋgá lóŋgómiŋgí

B H B H B H H H B H

"ma plume"

c. itómbo > ésitʃísombíso

H B H B H B

d. itómbo imbatʃɛ > ésitʃísombíso esimbísatʃíse

H B H B H B B B H B H B

"Le chapeau est déchiré"

Les exemples (b et d) présentent le cas d'un pont tonal qui marque la fin d'un mot en cours d'actualisation et le début du mot suivant. Ceci est interprété comme un indice d'accès lexical qui entretient ici une relation assez déterminante avec le lexique mental parce qu'à ce moment, il centralise l'attention des tous les percepts auditifs.

Outre ce qui vient d'être mentionné, on distingue aussi la catégorie tonale limitée. Ce sont des langues qui partagent partiellement ou totalement le même schème tonal et qui s'écartent de la première catégorie par la manière dont réagissent les tons en contexte syntaxique ou acoustico-phonétique. Le phénomène tonal le plus marquant qui est susceptible d'être interprété comme indice d'accès au lexique est le transfert de tons de la fin du mot précédant au début du mot suivant. Ceci peut être formulé comme dans les exemples suivants:

a.

t1	t2	t3	t4
|	|	|	|
S	S	S	S

b.

t1	t2	t3	t4	t5	t6	t7	t8
|	|	|	|#	|	|	|	|
S	S	S	S	S	S	S	S

devient:

t1	t2	t3	t4	t3	t4	t7	t8
|	|	|	|#	|	|	|	|
S	S	S	S	S	S	S	S

où S = unité porteuse de tons et t=un ton.

Lorsque deux dérivés se suivent, les derniers tons du premier mot se propagent au début du mot suivant. Concrètement, le dernier mot inverse son schème tonal. Le schème haut-bas (H-B) affecte plutôt les deux premières syllabes. C'est le phénomène observé dans L5 et L7, comme le montrent les formes reprises ci-dessous :

a. ɔtámbá > otambápa 'arbre' (L5)

b. ɔtámbá ɔnɛ́nɛ > otambápa ɔnɛnadʒɛ 'grand arbre'

c. ɔloko > olokádʒe 'cœur' (L7)

d. ɔloko ohánélá > olokádʒo ohanelɛdʒa 'mauvais cœur'

Les données qui précèdent montrent que les tons sont des faits prosodiques qui contribuent largement à l'identification lexicale; d'autant plus qu'ils font partie du rythme du langage qui est un phénomène inné mais différent en fonction de la langue et grâce auquel les enfants parviennent à réaliser leurs premières segmentations [Cut94]. Pour apprendre le français, les enfants se basent sur le rythme syllabique; pour l'anglais, ils tiennent compte du rythme accentuel ; enfin en japonais, ils se réfèrent à la more. En se fondant sur le fait qu'à moins de six mois, l'enfant est capable de rassembler les différences rythmiques en deux groupes, tons haut et bas [Dem77], on peut affirmer que ces faits prosodiques ont une réalité mentale. C'est une des raisons qui font poussent Jusczyk [Jus93] à déclarer que la structure prosodique est une dimension que l'enfant exploite pour accomplir la segmentation de la parole.

6. LA CATEGORIE IMPOSEE UNITONALE

Dans cette catégorie, tous les tons de la langue de départ sont réduits en une forme tonale simple; haute ou basse. Ces formes sont homotones comme dans L18 et 19, ainsi que le montrent les exemples suivants:

		L18	L19	
otámbá	>	mba:o	ta:tao	arbre
ákɔta otámbá	>	kɔ:a mbá:o	kɔ:kɔa tá:tao	il abat un arbre

Ces deux langues sont mutativo-hypocoristiques dans leur formation intentionnellement atone. La priorité prosodique semble se déplacer du rythme vers la durée parce que tous les paradigmes propres à ces langues doivent se conformer au gabarit prosodique suivant:

[XXX(x x) XX]

(où X = position squelettale pure; x = une position squelettale non indiquée pour L18)

Malgré son absence du mode de formation de ces deux langues dérivées, le ton haut (H) surgit quand il y a des paradigmes consécutifs dans un contexte syntactique. Cette émergence tonale au début du second paradigme est assez saillante au point qu'on peut dire qu'elle joue le rôle de démarcation entre les lexèmes consécutifs. Parce que ces deux langues sont prosodiques par leur mode de formation, le ton haut marque le début du second mot phonologique qui est trochaïque pour L18 et quadrimorique pour L19.

CONCLUSION

A la lumière de ce qui vient d'être dit, les tons ont un rôle important dans l'identification lexicale. Autant que d'autres unités pré-lexicales leur médiation fournit de précieux indices aux auditeurs pour marquer des repères dans le continuum du signal de la parole. Le phonème, le ton, la more, la syllabe ou le trait peuvent opérer conjointement dans l'identification lexicale. Il importe de préciser le caractère saillant de chacun par rapport au rythme de la langue concernée. Ces unités constituent des *representations intermédiaires* [Kol95] pour la reconnaissance des mots parlés tel dans le cas des dérivés initiatiques.

Cette recherche est subventionnée par la Convention ARC " Dynamique des systèmes phonologiques " 98-02, n° 226.

BIBLIOGRAPHIE

[Chu87] Church, K. (1987). 'Phonological parsing and lexical retrieval' Cognition, 25, 53-69.

[Con86] Content, A., Kolinsky, R., Morais, J. & Bertelson, P., (1986), " Phonetic segmentation in prereaders: Effets on corrective information ", Journal of Experiment Child Psychology, 42, pp 49-72.

[Cut94] Cutler, A. (1994), " Segmentations problems, rhythmic solutions ", Lingua, 92, pp 81-104.

[Cut88] Cutler, A. & Norris, D., (1988), " The role of strong syllables in segmentation for lexical access ", Journal of Experimental Psychology: Human perception and performance, 14, pp 113-121.

[Dem77] Demany, L., McKenzie, B. & Vurpillot, E., (1977), " Rhythm perception in early infancy ", Nature, 266, pp 718-719.

[Gro80] Grosjean, F., (1980) " Spoken word recognition processes and the gating paradigm ", Perception and Psychophysics, 28 (4), pp 267-283.

[Jus93] Jusczyk, P.W., (1993), " How word recognition evolves from infant speech recognition capacities " in Altman, G.T.M. & Shillcock, R.C. (eds) : Cognitive models of speech processing, The Sperlonga Meeting, II, pp 27-55, Cambridge MA, MIT Press.

[Kol95] Kolinsky, R., Morais, J. et Cluytens, M., (1995), " Intermediate representations in Spooken word recognition : Evidence from word illusions ", Journal of Memory and Language,34, pp 19-40.

[Meh68] Mehler, J., Segui, J. & Carey, P., (1968), " Tails of words: Monitoring ambiguity ", Journal of Verbal Learning and Verbal Behavior, 17, pp 29-35.

[Mor91] Morais, J., Castro, S.L. & Kolinsky, R., (1991), " La reconnaissance des mots chez les adultes illettrés " , dans Kolinsky, R., Morais, J. et Segui, J. (éds), La reconnaissance des mots dans les différentes modalités sensorielles, Paris, P.U.F.

[Rub68] Rubenstein, H. & Pollack, I., (1968), " Word predicability and intelligibility ", Journal of Verbal Learning and Verbal Behavior, 2, pp 147-158

[Tab93] Tabossi, P. (1993), Connections, competitions and cohorts : comments on the chapiters by Marslen-Wilson, Norris, Bard and Shillcock (eds) : Cognitive models of speech processing, The Sperlonga Meeting Hillsdade : Erlbaum, II, pp 277-294, Cambridge MA, MIT Press.

Auto-organisation induite par des fluctuations dans les systèmes phonologiques

S. C. Nicolis, J. L. Deneubourg*, A. Soquet°, D. Demolin°*

Université Libre de Bruxelles
*Centre d'Etudes de Phénomènes Nonlinéaires et des Systèmes Complexes, Bruxelles, Belgique.
° Laboratoire de Phonologie, Bruxelles, Belgique.
Tél.: +32 (2)650 5796. E-mail : snicolis@ulb.ac.be

ABSTRACT

Our purpose in this note is to study a system sharing some aspects of a phonological system. The problem of competition between a set of sounds initially present, one or some of which is used to designate an object, is explored and the idea that such a system may evolve via self-organized processes is developed. Self-organization is induced by processes of imitation and competition between individuals confronted to different options. Using Monte Carlo simulations, it is shown that in the simplest case of pair interactions and equal attractivities of the initial sounds the population evolves from an initial random state to a final organized one in which only one sound survives.

1. INTRODUCTION

Un problème important dans l'étude du langage et de la parole est d'expliquer la naissance et la dynamique de systèmes phonologiques. Le caractère inné de la capacité à émettre et reconnaître des sons et à les allouer à des objets a très longtemps été adopté par des chercheurs en phonologie [Cho68]. Pourtant, il paraît tentant d'explorer l'idée que ces systèmes, comme beaucoup d'autres qui sont biologiquement fondés, évoluent via des processus auto- organisés [Nic77]. Nous nous proposons ici d'aborder le problème sous cet angle avec des outils de modélisation et de simulation afin de voir comment une population d'individus parvient à finalement désigner un objet par un son. Notre principale thèse est que cette évolution se produit grâce à des « processus d'imitation » [Hey96] et de compétition entre individus confrontés à différentes options (informations).

Ce travail s'inscrit dans le cadre d'un modèle probabiliste simple qui cherche à mettre en évidence des mécanismes qui pourraient permettre de comprendre l'interaction entre deux ou plusieurs éléments d'un système phonologique. Les modalités d'interaction entre les éléments du système permettent de comprendre des aspects de la dynamique du système lui-même.

L'importance des mécanismes de l'imitation dans les processus cognitifs qui sont en jeu dans la dynamique du langage et plus particulièrement en phonologie a été récemment mise en évidence dans différents travaux ; notamment dans ceux de [Don91] où il est montré que l'imitation est une adaptation qui précède l'émergence du langage et de la phonologie, et dans ceux de [deB99] qui montre qu'en se fondant sur des principes très simples basés sur l'imitation, il est possible de simuler l'émergence des systèmes vocaliques. Le processus d'imitation auquel il est ici fait référence est très général et peut s'appliquer à de nombreux autres systèmes.

On adoptera le scénario le plus simple d'interactions par paires et d'options dont les attractivités sont à priori égales. On s'intéressera tout spécialement à l'influence de différents paramètres sur l'évolution, tels que le nombre de locuteurs ou des options à gérer, ainsi qu'aux aspects stochastiques inhérents à la dynamique [Gil92] [Van81].

Enfin, on s'attachera à dégager des perspectives expérimentales ouvertes par cette étude.

2. MODÈLE

Soit N une population d'individus capables d'émettre une série de sons, M pour désigner un même objet. Nous supposons que lorsque deux individus émettant les sons i et j se rencontrent, chacun peut, avec la même probabilité, convaincre l'autre que le son qu'il émet est le plus approprié pour désigner l'objet. Le processus cinétique correspondant peut être décrit comme

$$X_i + X_j \longrightarrow \begin{matrix} 2X_i \\ \text{ou} \\ 2X_j \end{matrix} \quad 2 < i, j < M \qquad (1)$$

Nous considérons le processus comme instantané et irréversible, ce qui ne veut pas dire que l'individu convaincu ne peut plus changer d'avis suite à une autre rencontre. A partir du schéma (1) on est tenté d'écrire un système d'équations différentielles décrivant la variation du nombre d'individus moyen $\bar{X}_i$ en fonction du temps, par analogie avec les équations de bilan de cinétique dynamique. Plus précisément, posons

$$\frac{dX_i}{dt} = \nu_{j\to i} - \nu_{i\to j} \qquad (2)$$

où $\nu_{j\to i}$ est la vitesse du processus de conversion de j en i et $\nu_{i\to j}$ celle de i en j ($j \neq i$). On s'attend à ce que chacune de ces vitesses soit donnée par la fréquence de rencontres de la paire i, j qui, en première approximation, devrait être proportionnelle au produit X_iX_j divisé par une facteur 2 afin d'éviter de compter la même paire deux fois. On est aussi amené à écrire:

$$\nu_{j\to i} = \frac{1}{2} X_i X_j$$
$$\nu_{i\to j} = \frac{1}{2} X_i X_j \qquad (3)$$

soit

$$\frac{dX_i}{dt} = \frac{1}{2} X_i X_j - \frac{1}{2} X_i X_j = 0 \qquad (4)$$

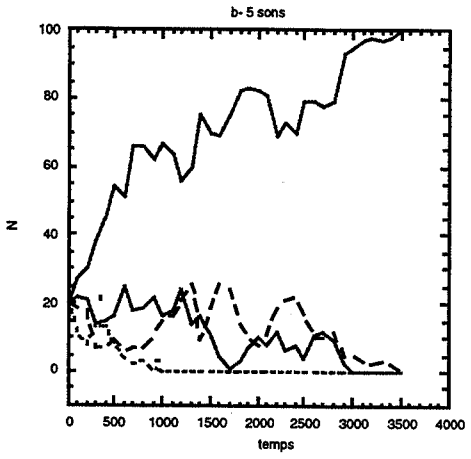
Nous arrivons aussi à la conclusion à première vue surprenante qu'en valeur moyenne aucune évolution ne peut apparaître. Cette conclusion reflète, en fait, la mise en échec d'un raisonnement basé sur les moyennes.
En effet, comme les rencontres entre deux types d'individus sont purement aléatoires en absence d'un « biais » favorisant l'une des options (eq. (1)), on peut s'attendre à ce que dans une population de taille finie, des évènements tels qu'une succession de rencontres où un individu émettant un son particulier impose successivement sa volonté à tous les partenaires rencontrés puisse se produire avec une probabilité non-nulle. A partir du moment où un tel événement se produit le système atteint un « point de non retour » suite à l'extinction de toutes les populations prononçant les sons non- retenus. Il est donc essentiel d'aller au delà du raisonnement macroscopique basé sur les eqs. (2)- (4) et d'incorporer explicitement la dynamique des fluctuations dans la description.

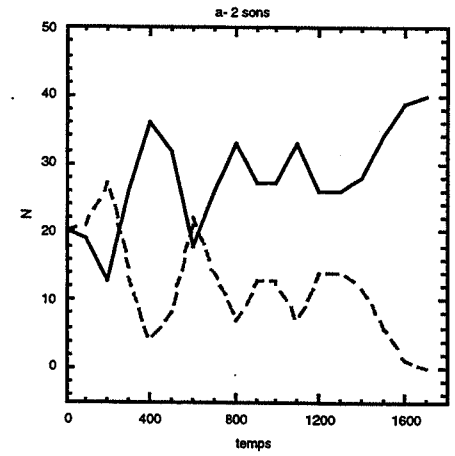
Pour réaliser cette description élargie nous faisons appel à la méthode de Monte Carlo qui consiste à effectuer une simulation directe du processus sous- jacent (plutôt que de résoudre des équations d'évolution des variables impliqués). Les principales étapes de cette démarche peuvent être résumés comme suit:

(i) A l'instant initial on désigne au hasard les individus (dont le nombre total N reste constant tout au long du processus) qui émettront un son particulier parmi les M sons en présence.

(ii) A chaque instant t=i deux individus sont tirés au hasard. Si ils font partie d'une sous- population d'individus émettant le même son ils sont remis dans la population et le processus recommence. S'ils font partie des sous- populations émettant des sons différents un des deux individus quitte sa sous –population d'origine et fait désormais partie de la sous-population de l'autre individu. Le choix de l'individu qui sera effectivement converti est aléatoire (probabilité 1/2). La composition de la population est réactualisée et le processus recommence étant entendu que les probabilités de tirage d'un individu émettant un son particulier sera affectée à la nouvelle composition.

Grâce à cette méthode on peut suivre à chaque pas de temps l'état de « conviction » des individus de la population. Les processus s'arrête si le système atteint un état « organisé » où tous les individus émettent un même son pour désigner l'objet.





Figures 1 : Évolution en fonction du temps du nombre d'individus prononçant le même son en désignant un objet, pour un nombre de sons égal à 2 (1a) et 5 (1b). Le nombre d' individus initial prononçant chaque son est égal à 20.

3. RÉSULTATS

Les Figures 1a et 1b résument des trajectoires stochastiques issues d'une telle simulation. On constate que, pour différents nombre de sons, et grâce aux fluctuations, le système évolue vers un état homogène, où tous les individus adoptent en définitive le même son pour désigner l'objet.

Il est intéressant de savoir quel est le temps moyen, pour un grand nombre de réalisations, nécessaire pour arriver à l'état homogène. Différentes simulations ont été entreprises où nous avons systématiquement pris un nombre de sons variant de 2 à 10 et des nombres d'individus émettant chaque son allant jusque 30.

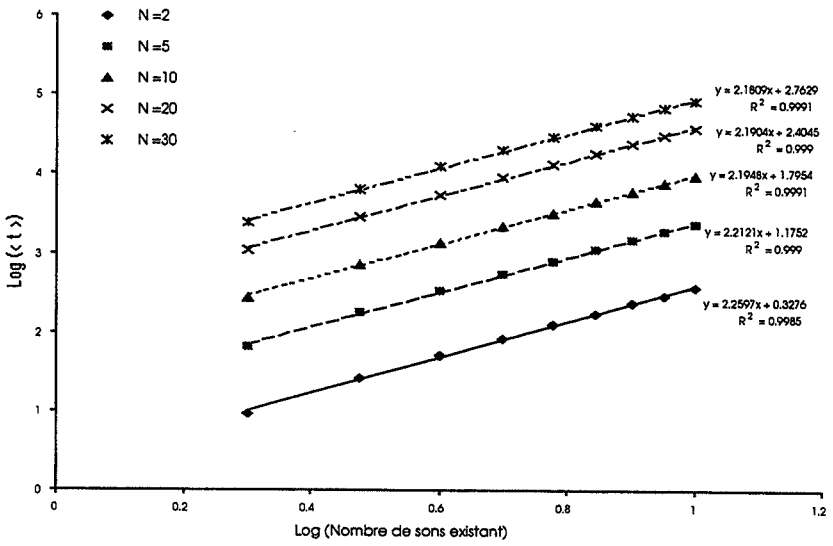


Figure 2 : Evolution (en échelle logarithmique) du temps moyen en fonction du nombre de sons pour différentes populations initiales émettant chaque son.

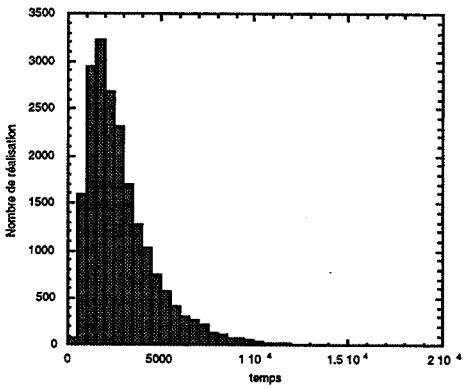


Figure 3 : Distribution des temps moyens pour arriver à 100 % de la population émettant le même son pour désigner un objet dans une situation où il y a initialement 3 sons, chacun ayant 20 individus pouvant émettre le son.

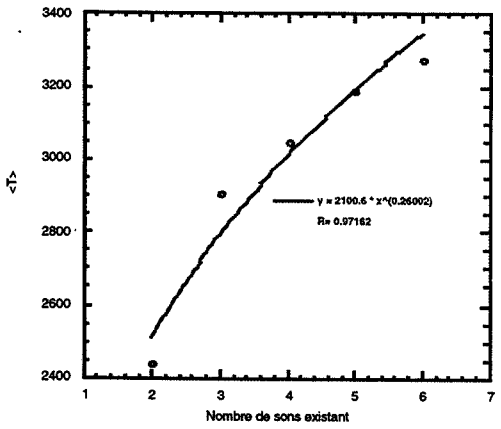


Figure 4 : Evolution du temps moyen pour arriver à 100 % d'individus prononçant le même son pour un même objet en fonction du nombre de sons présent initialement. La population totale est dans chaque cas égale à 60 et est divisé en sous populations initiales.

Nous regardons l'évolution du temps moyen pour arriver à 100% de la population qui émet un même son pour désigner un objet en fonction du nombre de sons, pour différents nombre d'individus émettant chaque sons. Nous constatons une tendance de croissance marquée.

Afin de préciser la loi suivie en fonction du nombre de son, nous avons pris une échelle logarithmique et nous avons dessiné une courbe de tendance linéaire. Ainsi que nous voyons sur la Figure 2, le temps moyen évolue comme une loi en puissance. Ceci reflète le fait que, contrairement à une situation régie par une loi exponentielle, des fluctuations importantes associées a des évènements à première vue exceptionnels peuvent se produire avec une probabilité appréciable. La Figure 3 confirme cette idée en montrant l'histogramme des temps pour arriver à 100 % de la population émettant un même son pour un objet pour les différentes réalisations. Nous voyons effectivement une très grande variabilité de ce temps, traduite par l'apparition d'une longue queue dans la distribution.

D'autre part, lorsque nous prenons une population constante et que nous divisons celle-ci en sous populations pouvant émettre initialement une série de sons, nous constatons que l'évolution du temps moyen pour arriver à 100% d'individus prononçant un même son pour un objet croît également. La Figure 4 (cercles ouverts) décrit ce genre de dynamique où la population totale est fixée à 60, le nombre de sons variant de 2 à 6. Les sous populations initiales sont alors, respectivement

pour chaque nombre de sons, égales à 30, 20, 15 et 10. La ligne pleine représente un ajustement en loi de puissance.

Nous voyons que l'exposant est peu élevé (0.26). Nous en concluons que l'effet du nombre d'individus sur les temps moyens est assez négligeable.

4. DISCUSSION

Nous nous sommes attachés dans ce travail à comprendre comment un système partageant certaines propriétés essentielles d'un système phonologique peut émerger à partir d'un état initialement non structuré. Notre étude s'est limitée à un modèle minimal dans lequel il n'y a qu'un objet à designer et les interactions entre individus sont entièrement aléatoires. Il serait important de généraliser ce modèle pour incorporer des éléments plus réalistes, comme par exemple un nombre d'objets plus grand que un à identifier par des sons ou des probabilités qui changent selon qu'un son est d'un type ou d'un autre. Des résultats préliminaires montrent que dans ce dernier cas on est conduit à une situation où il y a coexistence entre différents sons.

Il est intéressant de constater qu'à partir de règles simples, on retrouve des comportements propres aux systèmes plus complexes comme les systèmes phonologiques ou le langage, qui sont des systèmes de communication biologiquement fondés [Ede89]
Enfin, un test expérimental des principales conclusions de la présente étude peut être envisagé. Il consisterait à induire une compétition entre différentes informations du point de vue de leurs qualité (valeur sélective) au sein d'une population. En suivant la dynamique des choix collectifs qui seront opérés pour la population, on sera en mesure de construire des schémas analogues à ceux considérés ici et de comparer leurs comportements expérimentaux à ceux que nous avons mis en évidence.

REMERCIEMENTS

Cette recherche est subventionnée par la convention ARC "Dynamique des systèmes phonologiques", 98-02 n°226.

BIBLIOGRAPHIE

[Cho68] Chomsky M. & Halle M. (1968). The sound Pattern of English. New York. Harper and Row.

[Nic77] Nicolis G. & Prigogine I. (1977). Self-organization in Nonequilibrium system. New York: Wiley.

[Hey96] Heyes C. N. & Galef B. G. (1996). Social Learning in Animals: The Roots of Culture. Academic Press.

[Don91] Donald M. (1991). The Origin of Modern Mind. Cambridge. Harvard University Press.

[deB99] DeBoer B. (1999). Self-Organisation in Vowel Systems. Thèse de Doctorat. Vrije Universiteit Brussel.

[Gil92] Gillespie D. T. (1992). Markov Processes, Academic Press San Diego.

[Van81] Van Kampen N. G. (1981). Stochastic Processes in Physics and Chemistry. Amsterdam: North-Holland.

[Ede89] Edelman E. (1989). Bright Air Brillant Fire. New-York. Basic Books.

Modélisation de la prosodie par formes globales : amont ou aval de la phonologie tonale ? L'exemple d'un modèle développé à l'ICP

Véronique Aubergé

Institut de la Communication Parlée, UMR CNRS 5009, Université Stendhal/INPG, Grenoble
Tél. : +33 (0)476 82 41 97 - Fax : +33 (0)476 82 43 35
Mél : auberge@icp.inpg.fr

ABSTRACT

Is proposed here, a point of view consisting in presenting prosody as an emergent form perceived within a Gestalt type of processing. In this perspective, tonal phonology may be interpreted as a bottom-up sub-processing approach, and global forms modelling as a top-down approach. Although the model presented here is not described in detail, it may serve as an example to illustrate our hypotheses. Further, some significant perception results are recalled.

1. INTRODUCTION

Lorsque l'on adopte l'hypothèse d'un traitement globaliste de la prosodie, et que l'on se pose la question piège de la comparaison des approches phonologiques, phonétiques et linguistiques de la prosodie, on peut se retrouver devant un paradoxe étrange.

La tentation est grande de présenter la tradition de la phonologie tonale comme séparant les représentations symboliques de la substance prosodique à un niveau précoce (l'abstraction du ton est du domaine de la syllabe ou de la more), alors que parmi les approches phonétiques, certaines renvoient implicitement ce niveau à des sous-traitements d'un processus cognitif, avec ou sans abstraction phonologique. Celui-ci activerait tardivement et holistiquement, et surtout directement, les fonctions linguistiques.

Nous n'aurons pas du tout la prétention ici de dresser un état de l'art de ces approches phonologiques vs. phonétiques. Par contre, nous essayons de montrer qu'il existe une voie triviale, explicative sur le plan procédural cognitif et non pas descriptive, par laquelle la prosodie fonctionnerait par un processus *Gestaltiste* sur des formes globales *émergentes* de ces unités sur lesquelles repose l'abstraction en phonologie tonale.

Nous illustrerons cette alternative théorique par les conclusions que nous tirons des principaux développements autour d'un modèle de génération de la prosodie dont plusieurs implémentations successives, différemment enrichies par leurs auteurs, ont été développées à l'ICP à partir de 1991, et sur lesquelles nous ne reviendrons pas ici. Le but de cette modélisation – même si des simulations ont débouché sur des synthétiseurs – n'est pas de décrire les représentations physiques de la prosodie, mais de proposer des indices sur les traitements cognitifs qui participent au sens. Selon cette définition, il s'agit d'un modèle phonologique. Nous nous sommes attachés à décrire la morphologie prosodique, et ses contraintes internes, comme une représentation guidée par les fonctions qu'elle remplit dans le système communicatif (cf. figure 1). Cette démarche résolument descendante, est installée sur une méthodologie hypothético-inductive classique de la phonétique.

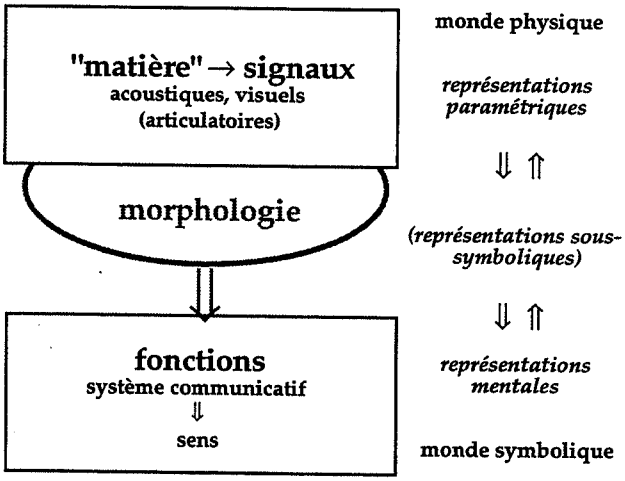


Figure 1 : démarche ascendante vs. descendante.

2. PHONOLOGIE TONALE VS. PHONÉTIQUE PROSODIQUE : Contradiction ou continuité ?

Le paradoxe qui nous préoccupe a été initié par un constat simple : caricaturalement on peut dire que la phonologie tonale manipule une représentation symbolique, le ton, dont la nature et l'universalité varie selon les théories [TOBI, Gussenhoven, Mertens, Hirst...]. Ses corrélats perceptifs sont établis plus sur des critères psycho-acoustiques que fonctionnellement linguistiques : c'est le même objet – la même capacité perceptive basique – qui est utilisé en accès lexical des langues à tons. Cette unité se dégage à un niveau de granularité de l'ordre de la syllabe, et se localise donc par rapport à la chaîne phonémique (ce qui permet ainsi d'articuler naturellement le lien entre segmental et supra-segmental selon la terminologie américaine). Les discussions sur la nature des unités candidates (syllabe, syllabe accentuée, more, pied, groupe inter-centre-perceptif...) à l'intérieur des débats, par exemple de la théorie métrique, ne changent en rien le degré de granularité de l'objet symbolique désigné et surtout son ancrage à la chaîne phonémique.

L'activation des fonctions linguistiques est ainsi en général une grammaire plus au moins complexe sur ces tons qui relie les structures tonales aux autres structures

linguistiques de l'énoncé. Cependant, nous n'avons pas su trouver dans cette littérature des arguments formels ou linguistiques qui nierait la notion de contours en tant qu'objets prosodiques, elle est même, dans certaines théories, comme celle de Hirst, calculée explicitement à partir des tons.

Dans les modèles issus du domaine de la phonétique, pour le français en particulier [Lac99], on rencontre de nombreux modèles par contours. Il nous semble important de les séparer en deux catégories : (type 1) ceux qui représentent les énoncés prosodiques par une combinaison de contours élémentaires dont la nature et la valeur peuvent ou non varier selon les langues (voir par exemple le modèle de l'IPO où les contours sont définis par des filtres perceptifs, ou ceux de Vaissière qui s'appuie d'abord par l'analyse acoustique ; le domaine du contour n'est pas forcément en liaison forte avec la syllabe et sa localisation n'est pas forcément asservie aux marquages linguistiques (type 2) les modèles pour lesquels les contours sont globaux au sens où ils ne sont ni décomposables ni "composés" pour devenir des contours supérieurs. On y trouve en particulier les modèles superpositionnels dont le plus connu est celui de Fujisaki. La démarche de celui-ci est avant tout ascendante et guidée par des contraintes intrinsèques (physiologiques) aux formes. Ce modèle, très souvent repris, a été associé à un guidage linguistique mais reste avant tout un modèle ascendant. D'autres modèles, non superpositionnels, associent directement une valeur à l'unité linguistique sur laquelle ils s'étendent. Ce sont typiquement les patrons de base de Delattre ou les clichés mélodiques de Fonagy. C'est dans cette dernière catégorie que nous situons notre modèle, pourtant superpositionnel, défini à partir des fonctions communicatives qu'il value, et décliné sur une hiérarchie de domaines. Si nous ajoutons une notion d'ordre morphogénétique à l'élaboration des contours, celle de l'émergence de Morgan, alors nous pourrions ordonner (plutôt qu'opposer) les approches phonologiques, phonétiques et "linguistiques", en utilisant la métaphore du pont (cf. figure 2), reprise de Searle par Moeschler.

Ainsi pour la phonologie tonale, "l'essence" prosodique ne serait pas dans les propriétés abstraites du pont mais dans celles des pierres, la caractérisation des pierres maîtresses et leur architecture. L'accès aux fonctions du pont est indirect. L'opposition de contours ne paraît pas immédiate, ni par paire minimale sur le système, ni par "traits tonals" sur les contours. Pour la première catégorie des modèles phonétiques, le pont est un assemblage (par contraintes ou par grammaires) de contours élémentaires, distribués sur des zones définies morphologiquement comme nécessairement décrites. Pour la deuxième c'est le pont lui-même qui active les fonctions. Les modèles linguistiques se réservant la description de l'usage des fonctions dans le système communicatif [Cou96] : faire passer les voitures sur le pont ! Même si notre modèle n'introduit par un prédécoupage du puzzle prosodique en tons en tant que sous-unités de calcul, mais utilise bien directement des contours phonétiques, nous voulons bien faire l'hypothèse qu'il existerait un processus d'émergence entre les tons et le contour : ainsi "l'essence" prosodique, l'objet qui fait sens, est la forme du contour elle-même. Les propriétés intrinsèques à la forme prosodique qui permettrait une perception par Gestalt utiliserait donc des capacités de production et de perception tonale, non pas comme représentations mentales, mais comme prédécoupage morphologique. Ce qui remet en cause la nature strictement symbolique du ton, mais qui permet d'expliquer comment pourrait fonctionner une construction où l'on s'attacherait à décrire les pièces les plus pertinentes du puzzle, "les oreilles de l'âne" qui permettent avec quelques morceaux bien choisis de l'image d'identifier globalement le bon équidé. Des exemples et contre-exemples sont donnés expérimentalement dans des systèmes de génération de la prosodie de type CHATR lorsque les "oreilles d'âne" sont effacées et empêche le décodage perceptif ou au contraire permettent à un contexte prosodique mal formé de résister au décodage.

Il existe quelques pistes objectives qui supposent que la construction de "l'image prosodique" par des pièces bien choisies combinées (type 1) ou surtout par une grammaire sur une abstraction de ces pièces (la phonologie tonale) est cognitivement artefactuelle, même si elle peut être descriptive : dans une perception globale, la forme n'est pas le résultat d'une grammaire calculable sur des unités de la forme. Si c'était le cas, l'identification de la forme (i.e. sa valeur fonctionnelle) interviendrait à la fin de l'énoncé sur lequel porte la forme. Or, déjà en 83, Grosjean a montré que les auditeurs sont capables de prédire, sur des critères rythmiques, la longueur manquante d'un énoncé tronqué. Thorsen [Tho80], van Heuven et al [Heu97], Grépillat et Aubergé [Aub97] ont pu mettre en évidence que les auditeurs identifient précocement la valeur des modalités ou attitudes véhiculées par les contours (par des tests de dévoilement progressif, Grépillat montre en particulier qu'en français, dès la deuxième syllabe d'énoncés à cinq syllabes, les auditeurs identifient la bonne attitude parmi six).

Pour résumer, si l'on pose ces deux hypothèses d'émergence morphologique et de perception globale :

- on peut situer la phonologie tonale dans une démarche ascendante, ancrée dans la chaîne segmentale puis motivée par les fonctions;
- la phonétique des contours globaux se situe dans une démarche descendante, motivée d'abord par l'activation des fonctions et qui déboucherait, si l'on conserve une approche classique, sur des prosodèmes ou intonèmes (lorsque l'on s'intéresse aux fonctions restreintes de l'intonation) qui organisent la prosodie en classes de contours et qui devraient donc pouvoir devenir des sujets de perception catégorielle.

3. NOTRE MODÈLE

Ce modèle s'inscrit dans l'hypothèse large que les fonctions remplies par la prosodie (linguis-

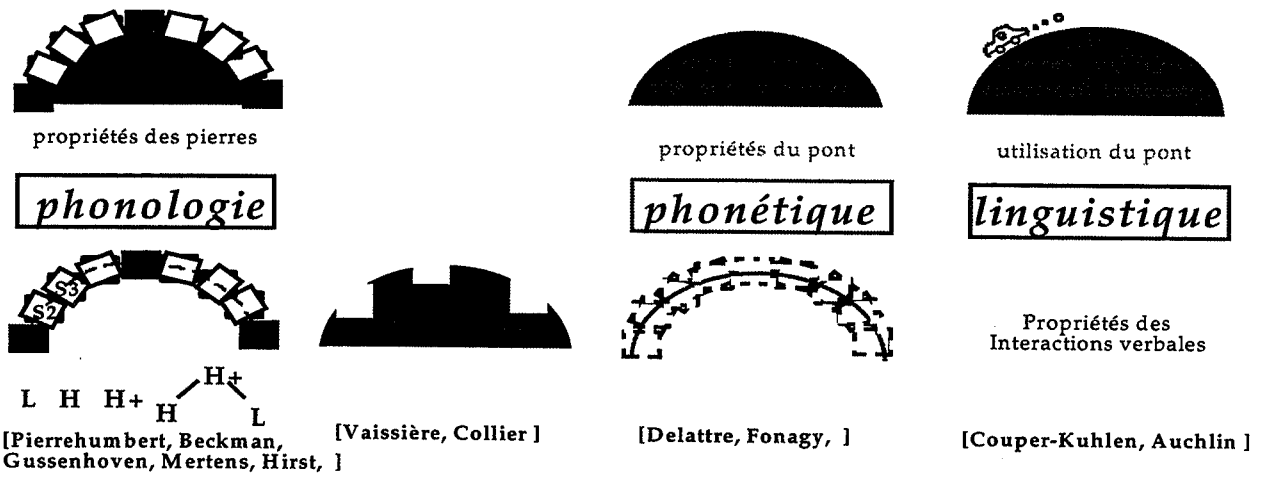


Figure 2 : la morphologie prosodie : un traitement global sur des contours émergents.

tiques,.pragmatiques, émotionnelles) sont globales au système de communication et réparties *interactivement* entre les différents agents du système, dont la prosodie (cf. figure 3). Cette hypothèse n'est pas modulaire au sens de Fodor puisqu'il n'y a pas de planification par un agent central, mais au sens des principes de "l'auto-organisation dirigée" proposée dans les modèles de l'intelligence du vivant.

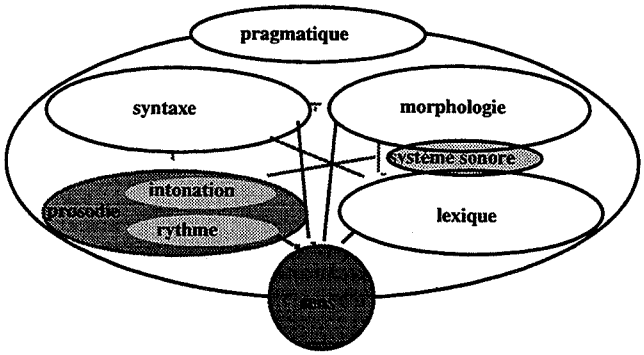


Figure 3 : les fonctions du système : réparties sur des agents morphologiques en coopération interactive

Elle nous renvoie (1) au niveau morphogénétique à l'autonomie/intégration prosodie/système sonore (2) au niveau cognitif à l'autonomie/coopération entre la prosodie et les autres agents. Les observations développementales (rassemblées par Demuth [Dem97]) de "boostrapping" entre prosodie et syntaxe corroborent cette hypothèse.

Nous avons choisi de modéliser la prosodie en l'isolant dans la réalisation de plusieurs de ses fonctions (démarcation, modalisation, focalisation ; expression des attitudes ; expression des émotions). Le domaine de variation morphologique de la réalisation de chaque fonction est étudiée séparément, c'est-à-dire en bloquant la variation des autres fonctions.

La méthodologie que nous suivons est toujours la même. (a) La première étape est de définir quelles valeurs de la fonction sont réalisées par la prosodie. (b) Ensuite il s'agit de fabriquer un corpus qui représente un large domaine de variation avec une densité la plus lourde possible [Aub92]. (c) La réalisation quantitative et qualitative des valeurs fonctionnelles est vérifiée perceptivement [Aub97 ; Ril99]. (d) De ce corpus est extrait une quantité valide de variantes de chaque classe de contours, une classe étant désignée par le fait qu'elle active une valeur de la fonction étudiée. Ces variantes seront représentées dans le modèle par un élément statistique (le contour-moyen dans la première implémentation [Aub92], le prototype calculé par un réseau connexionniste dans la seconde [Mor98]). (e) Le modèle est activé pour produire des énoncés synthétiques dont on évalue les capacités à reproduire les mêmes performances perceptives que le corpus orignal [Mor99; Ril99].

Il s'agit bien là d'une démarche descendante, même si elle est vérifiée sur les données, et non pas au contraire dirigée par les données. Si le corpus est un outil obligatoire du modèle, c'est que les éléments manipulés ne sont pas symboliques : ce sont des prototypes acoustiques, multi-paramétriques (non pas réduits à la simple évolution du fondamental) qui pourraient être dans le même principe audiovisuels et/ou articulatoires : ce modèle n'a pas besoin d'une étape d'abstraction, phonologique, des contours. Pourtant, les objets modélisés sont intrinsèquement des représentants centraux de classes qui s'identifient par les valeurs des fonctions et s'opposent perceptivement. Pour rester cependant dans une approche phonologique classique, il faudrait prouver réellement un processus de perception catégorielle et définir sur un plan phonétique comment ces formes sont "de bonnes formes" selon des contraintes de plus bas niveau (à la suite par exemple des travaux de Fujisaki ou Maeda).

3.1. La fonction de démarcation

La première fonction que nous avons étudiée, à la suite d'une majorité d'auteurs, est la démarcation (segmentation/hiérarchisation des énoncés). Les liens que la prosodie (l'intonation) entretient avec la morpho-syntaxe sont, rappelons-le, résultants de la coopération interactive entre ces deux agents, prosodie et morpho-syntaxe. Un corpus (de phrases isolées, lues) a donc été construit autour de ces *rendez-vous* entre les deux structures. Une analyse statistique a filtré les contours organisés autour des rendez-vous effectivement réalisés dans le corpus [Aub92]. Il a pu ainsi être mis en évidence que pour cette fonction, les réalisations prosodiques sont structurées hiérarchiquement (phrase, clause, groupe,

sous-groupe). Chaque niveau renvoie à un espace de contours organisés en classes, l'étiquette "phonologique" d'une classe étant sa valeur fonctionnelle pour ce niveau hiérarchique, et le domaine de répartition du contour étant le domaine de l'unité linguistique qui entretient un rendez-vous avec la morpho-syntaxe. Ainsi les valeurs fonctionnelles du niveau phrase sont les modalités, nous verrons au § 3.2 que sur ce même domaine portent également des valeurs attitudinales. La construction d'un patron prosodique est une superposition des contours de chacun des niveaux impliqués. Il est important de noter que dans cette étude, (1) la morphologie des contours d'un niveau est *indépendante* des autres niveaux ; par contre lors de l'opération de superposition qui consiste à ajouter des contours *portés* sur un contours *porteur* (chaque porté pouvant devenir porteur de contours du niveau inférieur) l'interaction d'un niveau sur l'autre devrait être intégrée dans le modèle (ce qui n'est pas encore le cas) afin de rendre compte de contraintes morphologiques mais également stratégiques dans la superposition (chevauchement, élasticité et "force" des contours) (2) que des rendez-vous semblent obligatoires, d'autres jamais réalisés, d'autres encore facultatifs, et que certains rendez-vous soient déplaçables [Cam93]. Une étude perceptive menée par Morlec [Mor98] sur des stimuli synthétiques construits avec quelques incohérences majeures entre les valeurs démarcatives de la prosodie et de la syntaxe sur des domaines cohérents, confirme l'analyse acoustique. Rilliard [Ril99] a réalisé une étude plus fine sur l'identification uniquement prosodique des valeurs de démarcation avec de la parole délexicalisée. Ce travail a permis de commencer à construire une sorte de grille d'évaluation de l'intelligibilité démarcative de la prosodie (en parallèle pour un corpus donné, et pour un modèle de synthèse donné). Ces résultats vont dans le sens d'une autonomie cognitive de la prosodie (puisque les auditeurs sont capables d'identification avec une prosodie isolée artefactuellement) et permet de tracer, pour un locuteur dans une situation figée, une stratégie de coopération entre prosodie et morpho-syntaxe.

3.2. La fonction attitude

Morlec [Mor98] dans son implémentation du modèle par un réseau connexionniste récurrent a réalisé un classifieur efficace capable de généraliser les mouvements prosodiques sur la longueur des énoncés. Il s'est essentiellement attaché au domaine de la phrase et à la fonction pragmatique des attitudes du locuteur. Afin de constituer un corpus à partir duquel le réseau pouvait classifier les contours porteurs du niveau phrase sans le "bruit" des niveaux inférieurs impliqués dans la fonction démarcative, le corpus a été constitué de monomots énoncés dans 6 attitudes.

3. 3. La fonction focalisation

Cette étude en cours n'a pas encore abouti sur un modèle de génération. Cependant nous voulons insister sur le fait que l'intégration morphologique de la réalisation de cette fonction dans le modèle de superposition ne s'énonce pas en terme de local vs. global. Il s'agit là pour nous d'une fausse opposition, puisque le modèle de superposition consiste justement à intégrer des formes indépendantes, liées à un domaine de répartition, et que ce domaine n'est pas plus local lorsqu'il s'agit d'une focalisation lexicale que lorsqu'il s'agit d'une valeur sur une clause syntaxique. Nous espérons de ce travail qu'il nous aide justement à formaliser ce "tuning" qui, dans notre modèle de superposition, devrait permettre à un contour porteur d'intégrer un contour porté avec des degrés de liberté imposés d'abord par les motivations du système de communication (force de la focalisation, hyper-intelligibilité...), et négociés ensuite avec les contraintes de production/perception.

4. Conclusion

Dans une perspective de modélisation cognitive de la prosodie, un modèle "direct", comme celui que nous prposons ici, peut donner lieu à deux types de conclusion: soit l'étape de symbolisation tonale est un artefact (efficace) de calcul qui formalise l'articulation segmental/supra-segmental, soit elle appartient aux traitements cognitifs mais intervient alors en sous-traitement qui permettrait (en supposant que des attributs morphologiques restent associés aux symboles) l'intégration de la prosodie et du système phonémique. Dans les deux cas, toujours selon la mêmehypothèse, la phonologie tonale ne peut pas être considérée comme une démarche descendante, motivée par la participation de ces objets sonores au sens.

Bibliographie

[Lac99] Lacheret A. Beaugendre F. (1999) La prosodie du français, CNRS Ed.

[Cou96] Couper-Kuhlen E. Selting M (1996) Prosody in conversation, Cambridge Un Press.

[Tho80] Thorsen N. (1980) "A study of perception of sentence intonation-evidence from Danish," J. Acoust. Soc. Am. 67 (3), 1014-1030.

[Heu97] van Heuven V., Haan J, Janse E. & van der Torre E. (1997) Perceputual Identification of sentence type and the time-distribution of prosodic interrrogativity markers in Dutch, Intonation ESCA WS, 317-320, Athènes.

[Aub92] Aubergé V. (1992) Developing a structured lexicon for synthesis of prosody, in Talking machine, Bailly & Benoit Eds, Elsevier

[Ril99] Rilliard A. (1999), Prosody diagnostic using reiterant speech, ICPhS, 37-40, San Francisco.

[Mor98] Morlec Y. (1998) Génération multiparamétrique de la prosodie du français, thèse de l'INPG.

[Aub97] Aubergé V.,Grépillat & Rilliard A. (1997), Can we perceive attitudes before the end of sentences? A gating paradigm for prosodic contours, Eurospeech, Rhodes.

[Mor98] Morlec Y. Rilliard A. Bailly G. & Aubergé V. (1998) Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: methodology and first results, 1st LREC, Grenade.

[Cam93] Campbell N. (1993) Automatic Detection of prosodic boundaries in speech, Speech Communication, 13, 343-354

Reconnaissance et modèles de langage

Introduction de la Vitesse d'élocution dans un modèle de reconnaissance automatique de la parole

Abdellah Yousfi, Abdelouafi Meziane

Département de Mathématique, Faculté des sciences, Université Mohamed premier-Oujda, Maroc
(Tel : 50-06-01), (Fax : 50-06-03)
E-mail: yousfi.abdellah@sciences.univ-oujda.ac.ma
meziane@sciences.univ-oujda.ac.ma

Abstract

In this paper, we propose an improvement to the TLHMM (Two Level Hidden Markov Model) model centisecond [Mez 99] applied to the sound duration. Indeed the distributions of this parameter depend on the elocution velocity. An adaptation of the recognition processus or in a post-processing is needed. The first adaptation is studied while proposing a model of the elocution velocity based on filtres of Kalman. The second adaptation is based on the average syllabic duration. The experiments elaborated on a set of BDSONS show the interest of those approachs.

1. Définition du TLHMM centiseconde

Le modèle à deux niveaux (TLHMM: Two Level Hidden Markov Models) " centiseconde" s'inspire des HMM couramment utilisés en reconnaissance automatique de la parole. Il est organisé de manière hiérarchique à partir d'unités élémentaires. Au niveau syntaxique, la phrase est décrite sous la forme d'une concaténation des modèles de mots. Au niveau lexical chaque mot du vocabulaire est représenté par une séquence d'unités phonétiques et traité comme une concaténation des modèles acoustiques. Au niveau acoustico-phonétique, un modèle acoustique markovien est associé à chaque unité phonétique. Le modèle global est obtenu en compilant l'ensemble des modèles.
Dans l'approche segmental [Sua 94], le pré traitement acoustique du signal de parole, est segmental, chaque observation est obtenue sur des fenêtres de taille variable. Dans le cas du modèle TLHMM centiseconde, l'analyse acoustique est faite sur des trames de longeur fixe.
Le modèle de Markov caché à deux niveaux (TLHMM) centiseconde est défini à partir de trois processus stochastiques $(X_t,\ Y_t,\ t \geq 1)$ et
$(D_\tau,\ \tau \geq 1)$.
- $(X_t)_{t\geq 1}$ est un processus markovien caché d'ordre 1, à valeurs dans un ensemble fini d'états Q.
- $(Y_t)_{t\geq 1}$ est un processus observable représentant les observations acoustiques, à valeurs dans un ensemble mesurable Y.
- $(D_\tau)_{\tau\geq 1}$ est le processus prosodique. Étant donnée une suite d'états $(X_t)_{t\geq 1}$, correspondant à la modélisation d'une suite d'unités phonétiques $(\Phi_\tau)_{\tau\geq 1}$, D_τ correspond à la durée du son Φ_τ, durée de séjour dans la suite des états cachés extraits de $(X_t)_{t\geq 1}$, correspondant au son Φ_τ.
On pose :
- $(\wedge_\tau)_{1\leq\tau\leq\epsilon} = \{(\phi_{k_i}, \theta_i) \mid i = 1,..,\epsilon\}$ la suite phonétique correspondant à un chemin ou suite d'états $\xi_T = q_{i_1}, ..., q_{i_T}$ de longueur T,
- ϵ représente le nombre total d'unités phonétiques traversées lorsqu'on parcourt la suite d'états $(X_t)_{1\leq t\leq T}$.
- $\Phi_\tau = \phi_k$ à valeurs dans l'ensemble fini Σ des unités phonétiques élémentaires, $\Sigma = \{\phi_1, ..., \phi_k\}$, représente la τ^{eme} unité phonétique traversée lorsque nous empruntons la suite d'états $(X_t)_{t\geq 1}$ dans le sens des indices temporels croissants.
- $\Theta_\tau = \theta_\tau$ représente l'indice temporel du 1^{er} état de la suite $(X_t)_{t\geq 1}$ issu de la τ^{eme} unité. Cet indice correspond à un instant de changement de modèles acoustiques élémentaires.
- $(d_1, .., d_\epsilon)$ la suite de durées phonétiques déduites de $(\wedge_\tau)_{1\leq\tau\leq\epsilon}$,
Nous supposons de plus que la durée de séjour dans une unité phonétique ne dépend que de cette dernière :
$Pr(D_\tau = d_\tau \mid (\wedge_\theta)_{1\leq\theta\leq\epsilon}) = Pr(D_\tau = d_\tau \mid \Phi_\tau = \phi_k) = \varphi_k(d_\tau)$
Avec ces hypothèses la vraisemblance conjointe de la suite d'observations acoustiques $(y_1, .., y_T)$, et prosodiques est donnée par :

$$Pr(Y_1, ..., Y_T = y_T, D_1 = d_1, ..., D_\epsilon = d_\epsilon) = \sum_{\xi_T} \pi_{i_1} b_{i_1}(y_1) \times \prod_{n=2}^{\theta_2 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_1}(d_1) \times \prod_{n=\theta_2}^{\theta_3 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_2}(d_2) \times ... \times \prod_{n=\theta_\epsilon}^{T} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_\epsilon}(d_\epsilon)$$

En phase de reconnaissance, le meilleur chemin est donné par la formule :

$$\xi^* = arg\max_{q_{i_1},...,q_{i_T}} Pr(y_1, .., y_T, d_1, .., d_\epsilon, q_{i_1}, ..., q_{i_T})$$

La recherche du meilleur chemin s'inspire de la procédure de Viterbi [For 73].

2. Utilisation de la vitesse d'élocution en utilisant les filtres de Kalman(Modèle (1))

Plusieurs travaux réalisés ont montré que la vitesse d'élocution contribue de manière significative à la

variabilité du signal de parole. La majorité de ces travaux se sont interessés, uniquement, soit à l'analyse acoustique du signal de parole [Hug 72], [Lin 63], soit à faire des tests perceptifs [Mil 81], [Por 78], [Ver 78],[Pic 60], pour étudier cette variabilité.
Notre recherche a pour objectif l'étude de l'influence de la vitesse d'élocution sur la durée des sons, en introduisant ce facteur dans le modèle TLHMM centiseconde. Pour décrire cette influence, nous utilisons les filtres de Kalman, et nous retenons aussi le même modèle d'état que celui proposé par N.Suaudeau [Sua 94].
Considérons une prononciation constituée d'unités phonétiques $\phi_{k_1}, .., \phi_{k_\tau}, .., \phi_{k_\epsilon}$ de durées mesurées $g(1), .., g(\tau), .., g(\epsilon)$. Ce modèle s'écrit :

$$\begin{cases} r(\tau) = r(\tau-1) + \omega(\tau-1) \\ g(\tau) = \mu_{k_\tau} r(\tau) + v(\tau) \end{cases} \quad (1)$$

- $r(\tau)$: variable aléatoire du débit d'élocution, elle est supposée gaussienne. $r(\tau)$ est discretisée de façon qu'elle soit constante sur chaque unité phonétique ϕ_{k_τ}.
- $g(\tau)$: la durée observée pour la τ^{eme} unité phonétique ϕ_{k_τ}.
- μ_{k_τ}, σ_{k_τ}: sont la moyenne et la variance de la loi de durée, dans le cas standard, associées à ϕ_{k_τ}.
- $\omega(\tau)$, $v(\tau)$: deux bruits blancs gaussiens de moyennes nulles et de variances $Q(\tau)$, $R(\tau) = {\sigma_{k_\tau}}^2$.
- r_0 : variable aléatoire représentant la condition initiale, elle est de moyenne $r(0/0)$ et de variance $p(0/0)$ estimées sur l'ensemble d'apprentissage selon un critère de moindres carrés.
- Le bruit $w(\tau)$ représente la variabilité interne à la prononciation du débit, sa variance est supposée plus petite que la variance inter-prononciation $p(0/0)$. Nous supposons qu'elle est de la forme $Q(\tau) = p(0/0)/k$ avec $k \geq 1$
K sera fixé à l'aide des expériences faites sur l'ensemble d'apprentissage.
Pour estimer le facteur d'élocution $r(\tau)$ nous utilisons le filtre de Kalman [Bou88], ce filtre introduit les définitions suivantes :
$r(\tau/\tau) = E[r(\tau)/g(1), .., g(\tau)]$ représente l'estimation de la moyenne de $r(\tau)$ sachant $g(1), .., g(\tau)$.
$p(\tau/\tau) = E[(r(\tau) - r(\tau/\tau))^2/g(1), .., g(\tau)]$ est la variance de l'erreur de cette estimation.
En développant ces formules et en tenant compte de (1) nous obtenons les formules récurrentes :

$$\begin{cases} r(\tau/\tau-1) & = r(\tau-1/\tau-1) \\ p(\tau/\tau-1) & = p(\tau-1/\tau-1) + Q(\tau-1) \\ r(\tau/\tau) & = r(\phi_{\tau+1}) = r(\tau/\tau-1) + \\ & \quad G(\tau)[g(\tau) - \mu_{k_\tau} r(\tau/\tau-1)] \end{cases}$$

$$\begin{cases} G(\tau) & = \mu_{k_\tau} p(\tau/\tau-1)[\mu_{k_\tau}^2 p(\tau/\tau-1) + R(\tau)]^{-1} \\ p(\tau/\tau) & = p(\phi_{\tau+1}) = (1 - G(\tau)\mu_{k_\tau})^2 p(\tau/\tau-1) + \\ & \quad R(\tau)G(\tau)^2 \end{cases}$$

$G(\tau)$ est le gain du filtre de Kalman.

2.1 Vraisemblance d'une suite d'observations

Etant donnée une suite d'observations $(y_1, y_2, .., y_T)$ générée par le modèle TLHMM centiseconde lorsqu'on emprunte la suite phonétique $(\wedge_i)_{1 \leq i \leq \epsilon} = \{(\phi_{k_i}, \theta_i) \mid i = 1, .., \epsilon\}$; la vraisemblance de la suite d'observation $(y_1, .., y_T)$ en tenant compte de la vitesse d'élocution est donnée par la formule :

$$Pr(y_1, .., y_T, d_1, .., d_\epsilon) = \sum_{\xi_T} \pi_{i_1} b_{i_1}(y_1) \times \prod_{n=2}^{\theta_2 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_1}^{r(\phi_{k_1}), p(\phi_{k_1})}(d_1) \times \prod_{n=\theta_2}^{\theta_3 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_2}^{r(\phi_{k_2}), p(\phi_{k_2})}(d_2) \times ... \times \prod_{n=\theta_\epsilon}^{T} a_{i_{n-1} i_n} b_{i_n}(y_n) \times \varphi_{k_\epsilon}^{r(\phi_{k_\epsilon}), p(\phi_{k_\epsilon})}(d_\epsilon)$$

- ξ_T un chemin de longueur T correspondant à la suite phonétique $(\wedge_i)_{1 \leq i \leq \epsilon}$.
- $\varphi_k^{r(\phi_k), p(\phi_k)}(.)$ est la loi de la durée modélisant le temps de séjour dans l'unité ϕ_k en tenant compte de la vitesse d'élocution, elle est de moyenne $\mu_{\phi_k} r(\phi_k)$ et de variance $\mu_{\phi_k}^2 p(\phi_k) + R(\phi_k)$.
- θ_τ est l'indice temporel du 1^{er} état issu de la τ^{eme} unité phonétique ϕ_{k_τ}.
- d_τ le nombre d'observations émises dans l'unité phonétique ϕ_{k_τ}, il vérifie $d_\tau = \theta_{\tau+1} - \theta_\tau$.

2.2 Extension du modèle (1)

Pour prendre mieux en compte la réalité physique du signal Kay [Kay 93] a proposé un second modèle d'état non linéaire :

$$\begin{cases} r(\tau) = r(\tau-1) + w(\tau-1) \\ g(\tau) = \mu_{k_\tau} \exp(r(\tau)) + v(\tau) \end{cases} \quad (2)$$

On remarque que l'équation d'observation n'est plus linéaire ce qui nous oblige à développer un filtre de Kalman étendu, dit EKF (Extended Kalman Filter) [Kay 93], ce filtre consiste à linéariser le modèle d'état (2) autour d'un point de référence : la dernière moyenne estimée $r(\tau - 1/\tau - 1)$.
On pose $f = \exp(r(\tau - 1/\tau - 1))$
Si on fait un développement de Taylor autour de $r(\tau - 1/\tau - 1)$, l'équation (1) devient :

$$\begin{cases} r(\tau) = & r(\tau-1) + w(\tau-1) \\ g(\tau) = & \mu_{k_\tau} r(\tau) f + \mu_{k_\tau} f - \\ & \mu_{k_\tau} r(\tau - 1/\tau - 1) f + v(\tau) \end{cases} \quad (3)$$

Pour ce modèle les équations classiques de Kalman s'ecrivent :
$r(\tau/\tau) = r(\tau - 1/\tau - 1) + G(\tau)[g(\tau) - \mu_{k_\tau} f]$
$P(\tau/\tau - 1) = P(\tau - 1/\tau - 1) + Q(\tau - 1)$
$G(\tau) = \mu_{k_\tau} f P(\tau/\tau - 1)[\mu_{k_\tau}^2 f^2 P(\tau/\tau - 1) + \sigma_{k_\tau}^2]^{-1}$
$P(\tau/\tau) = [1 - G(\tau)\mu_{k_\tau} f]^2 P(\tau/\tau - 1) + (G(\tau)\sigma_{k_\tau})^2$

Les techniques de recherche du meilleur chemin ne changent pas, par contre la moyenne et la variance de la loi de durée sont ajustées suivant le modèle (3).

3. Utilisation de la vitesse d'élocution dans une phase de post-traitement

Dans la plupart des études phonétiques on retient, la syllabe comme unité fondamentale pour caractériser la vitesse d'élocution [Ros 81]. Ces études ont montré l'existence des liens, entre la durée des phonèmes et la durée globale de la prononciation dont ils sont extraits [Gog 93], et des liens entre la durée des voyelles et le nombre de syllabes du mot [Leh 73],[Kla 75]. En se basant sur ces études N.Suaudeau [Sua 94] a choisi la durée syllabique moyenne pour calculer la vitesse d'élocution, et un modèle de régression linéaire simple pour décrire l'influence de la vitesse sur la durée des phonèmes. Ce modèle ne sera utilisé qu'au cours d'un post-traitement. Dans notre travail nous avons gardé le même modèle pour voir son influence sur le modèle TLHMM centiseconde.

Définition :
Considérons une prononciation w constituée des unités phonétique $\phi_{k_1}, ..., \phi_{k_{\epsilon'}}$.
La durée syllabique moyenne *Syl* de w est une variable aléatoire qui se définit comme suit :

$$Syl = \frac{\sum_{\tau=1}^{\epsilon'} g(\tau)}{s}$$

$g(\tau)$: la durée observée de l'unité phonétique ϕ_{k_τ}.
s : le nombre de syllabes constituant la prononciation w.

Nous retenons le même modèle que celui proposé par N.Suaudeau [Sau 94], et qui décrit l'influence de la durée syllabique moyenne (vitesse d'élocution) sur la durée des sons. Ce modèle d'élocution est :

$$g(\tau) = \alpha_{k_\tau} Syl + \beta_{k_\tau} + v(\tau) \qquad \tau = 1, .., \epsilon \qquad (4)$$

$(\alpha_{k_\tau}, \beta_{k_\tau})$: les paramétres du modèle d'élocution caractèrisant l'unité phonétique ϕ_{k_τ}.
$v(\tau)$: bruit blanc de variance $R(\tau) = {\sigma_{k_\tau}}^2$ (cas standard).
Les paramètres $(\alpha_{k_\tau}, \beta_{k_\tau})$ sont estimés sur l'ensemble d'apprentissage suivant la méthode des moindres carrés pour caractériser l'unité phonétique ϕ_{k_τ}.

3.1 Utilisation du modèle d'élocution en post-traitement

Pour introduire le modèle d'élocution (4) dans le modèle TLHMM centiseconde, nous avons suivi les étapes suivantes :
1) Pour un mot prononcé w, la reconnaissance avec le modèle TLHMM centiseconde nous donne une solution w' constituée des unités phonétiques $\phi_{k_1}, .., \phi_{k_{\epsilon'}}$.
Si on note par s le nombre de syllabes qui compose w', alors il existe un i tel que le nombre de syllabes composant w est $s+i$ avec $i = -N, ..., N$ (N un entier à fixer sur les expériences en pratique on trouve N=1), on peut supposer que la durée syllabique moyenne a priori de w est :

$$S^i = \frac{\sum_{\tau=1}^{\epsilon'} g(\tau)}{s+i}$$

2) Pour chaque valeur de la vitesse d'élocution S^i, on ajuste les paramètres de la loi de durée suivant le modèle d'état :

$$g(\tau) = \overline{\alpha}_{k_\tau} S^i + \overline{\beta}_{k_\tau} + v(\tau)$$

$\overline{\alpha}_{k_\tau}$,$\overline{\beta}_{k_\tau}$ sont les estimateurs de α_{k_τ}, β_{k_τ}.
Les nouvelles paramètres de la loi de durée ajustés sont :

$$\begin{cases} \overline{\mu}^i_{k_\tau} = \overline{\alpha}_{k_\tau} S^i + \overline{\beta}_{k_\tau} \\ \overline{\sigma}^i_{k_\tau} = \sigma_{k_\tau} \end{cases} \qquad (5)$$

3) Après avoir ajusté les paramètres du modèle TLHMM centiseconde, une reconnaissance par ces nouveaux paramètres est effectuée. Pour un mot w on a au plus $2N+1$ solutions on les note w'_i, $i = -N, .., N$ où chacune de ces solutions w'_i est associée à une durée syllabique moyenne S^i.
4) Pour ces 2N+1 solutions deux cas existent :
a) Toutes les solutions sont identiques on ne fait rien.
b) Il existe des solutions identiques, et d'autres différentes. Parmi les solutions identiques, nous retenons celle qui a la durée syllabique moyenne mesurée, la plus proche de la valeur S^i supposée a priori. Ensuite nous sélectionnons parmi cet ensemble de solutions restreint (ne contenant que des propositions différentes) celle qui vérifie le meilleur score.
Dans le cas où les solutions sont toutes différentes nous appliquons directement les critères de sélections. Ces critères sont caractérisés par les trois scores classiques (score-acoustique, score-chemin, score-durée). Nous avons remarqué que ces scores ne reflètent pas la confiance qui peut être accordée à une solution, pour cela nous avons proposé un nouveau score (score-rev) qui combine entre ces trois scores.

score-rev=score-dur si l'une des solutions w'_i verifie Delta-dur < Delta-acou < Delta-chemin

score-rev= score-chemin si l'une des solutions w'_i verifie Delta-dur > Delta-acou > Delta-chemin

score-rev= score-acous sinon

où :

$$\text{Delta-acous} = \frac{\text{score-acous}(w'_i)\text{-score-acous(w')}}{\text{score-acous(w')}}$$

$$\text{Delta-dur} = \frac{\text{score-dur}(w'_i)\text{-score-dur(w')}}{\text{score-dur(w')}}$$

$$\text{Delta-chemin} = \frac{(\text{score-chemin}(w'_i)\text{-score-chemin(w')})}{\text{score-chemin(w')}}$$

score-acous(w') : le score acous associé à w'
score-dur(w') : le score durée associé à w'
score-chemin(w') : le score chemin associé à w'

4. Expérimentation

A) Données et analyse acoustique : Le vocabulaire se compose des nombres de 0 à 19 extrait de la base des données (BDSONS). Chaque nombre est prononcé une seule fois par 20 locuteurs pour former un ensemble d'apprentissage, et une autre fois par ces mêmes locuteurs pour former l'ensemble test.
L'analyse acoustique est faite sur des fenêtres de longueur fixe égale à 32ms pour déterminer un vecteur de 9 coefficients cepstraux (MFCC).
B) Modélisation : le réseau est construit de manière hiérarchique, et utilise comme unité élémentaire le pseudo-diphone, unité qui tient compte de la partie stable d'un phonème et de la transition entre phonèmes ; chaque unité phonétique correspond à un MMC élémentaire et les lois de durée liées à la vitesse sont portées par le pseudo-diphone.
Plusieurs tests ont été effectués en faisant varier k et N. Les meilleurs résultats sont obtenus lorsque k=70 et N=1.
Les résultats obtenus pour les trois modèles sont rapportés dans les deux tableaux suivants.
Tableau1 : Taux d'erreur sur l'ensemble d'apprentissage et de test pour les modèles (1) et (EKF)

	Apprentissage	Test
Sans la durée	8.61%	10.56 %
Avec la durée	3.06%	4.17%
Modèle(1)	3.89%	4.72 %
Modèle(EKF)	4.17%	4.44%

Tableau2 : Taux d'erreur sur l'ensemble d'apprentissage et de test pour le modèle (4)

	Apprentissage	Test
Score-acous	2.5%	3.61 %
Score-chemin	2.78%	3.89 %
Score-durée	2.78%	4.17 %
Score-rev	2.22%	3.89 %

d) Conclusion : Nous remarquons pour la première modélisation de la vitesse d'élocution basée sur les filtres de Kalman que les taux d'erreurs avec la durée des sons et la vitesse d'élocution sont presque les mêmes, car cela est du à la base de données, qui a servi de support à ces tests, composée de prononciations prononcées à un rythme normal. Pour le deuxième modèle, c'est à dire lorsque nous avons utilisé la durée syllabique moyenne comme facteur d'ajustement, le taux d'erreur sur l'ensemble test s'améliore. Nous espérons avoir des résultats meilleurs en développant des critères de sélections qui combinent mieux les différents scores. Les deux modèles sont validés sur un corpus formé de mots courts -ne contenant pas assez de syllabes et qui sont dites à un rythme normal- nous pensons qu'une application sur un vocabulaire se composant de mots contenant un nombre assez important de syllabes, et dites à des rythmes différents comme dans le cas de la parole continue, mettra en évidence et de manière significative l'intérêt de l'introduction de ce facteur dans les modèles de reconnaissance automatique de la parole.

Bibliographie

[Pic 60] Pickett, J. M. and Decker, L. R., Time factors in perception of a double consonant. Language and Speech, 1960, 3, 11-17.

[Lin 63] Lindblom, B., Spectrographic study of vowel reduction. JASA, 1963 , Vol 35, pp 1773-1781.

[Hug 72] Huggins, A. W., On the perception of temporel phenomena in speech. JASA, 1972, Vol 51, pp 1279-1290.

[Leh 73] Lehiste, I., Rytmic units and syntactic units in production and perception. JASA, 1973, Vol 54, pp 1228-1234.

[For 73] Fornay D. R., The Viterbi Algorithm, Proc. IEEE, vol. 61, n 3, mai 1973.

[Klatt 75] Klatt, D. H., Vowel lengthening is syntactically determined in a connected discourse. Journal of phonetics, 1975, 3, pp 129-140.

[Por 78] Port, R. F., Effects of word-internal versus word-external tempo on voicing boundary for medial stop closure. JASA, 1978, Vol 63, S20.

[Ver 78] Verbrugge R. R. and Isenberg, D., Syllable timing and vowel perception. JASA, 1978, Vol 63, S4.

[Mil 81] J. L. Miller, Effects of speaking rate on segmental distinctions in perspective on the study of speech. P. D. Eimas and J. L. Miller (eds), Lawrence Erlbaum Associates, Publishers, 1981, pp 39-74.

[Ros 81] Rossi., L'intonation : de l'acoustique à la sématique GALF Groupe de la communication parlée, 1981 . pp 41-53

[Bou 88] Bouleau N., Processus stochastiques et application.1988, HERMANN, pp29-34

[Kay 93] Kay S. M., Fondamentals of statistical signal processing. Editor : A.V oppenheim, Prentice-Hall, 1993

[Gog 93] Gog Y. and TREUNIET W., Duration of phones as function of utterance length and its use in automatic speech recognition. EUROSPEECH'93, Berlin, September 1993.

[Sua 94] Suaudeau N., Un modèle probabiliste pour intègrer la dimension temporelle dans un système de reconnaissance automatique de parole, thèse de 3°cycle , April 94.

[Mez 99] Meziane A., Jacob B., André-Obrecht R., Modélisation de la durée des sons dans un système de reconnaissance automatique de la parole, C. R. Acad. Sci. Paris, t.327, Série II b, p.379-382, 1999.

Apports d'une modélisation par réseaux de neurones multicadences à la reconnaissance de la parole

*Robert van Kommer** *et Beat Hirsbrunner*†

*SWISSCOM AG - CIT-CT-SPI - Güterstrasse 5 - CH-3050 Berne SUISSE
†Université de Fribourg - Groupe PAI - Chemin du Musée 3 - CH-1700 Fribourg SUISSE
Tél.: ++41 (0)31 342 64 57 - Fax: ++41 (0)31 342 29 53
Mél: Robert.vanKommer@swisscom.com - http://www.swisscom.com

Abstract

In voice-activated teleservices, speech recognition systems have a hard task to withstand the encountered adverse conditions. In order to address this robustness issue, this paper describes a new approach based on multi-tiered speech models. The main underlying hypothesis is that some model structures may perform better in a training-operating mismatch condition. Experimental results on connected digit recognition are presented in two different cases, (i) for a set of Hidden Markov Model recognition systems, and (ii) for the new system based on a *Multirate Neural Network*. According to the results, the robust behaviour of the second system looks promising and motivates further research towards modular spatiotemporal modeling of speech.

1. Introduction

Dans le cadre des services de télécommunications, la mise en fonction d'un système de reconnaissance automatique de la parole nécessite une attention toute particulière. Notamment, en ce qui concerne le manque de robustesse qui se manifeste lorsque les conditions d'utilisations diffèrent de celles de l'entraînement des modèles. La présente étude est guidée par l'hypothèse que certaines structures de modèle pourraient mieux résister à des différences entre les conditions d'entraînement et d'utilisations. Cette idée est soutenue par les deux suppositions suivantes: (i) plus la structure des modèles est proche de la représentation interne, meilleures seraient ses possibilités de résister à des variations inconnues; (ii) une étude sur l'intelligibilité [Gre97] suggère qu'une structure hiérarchique en multi-étages serait mieux adaptée qu'une purement séquentielle.

Dans l'objectif d'une vérification expérimentale de cette idée, une nouvelle architecture constituée d'un réseau de neurones multicadence est réalisée. La robustesse du nouveau système est mesurée et comparée avec des systèmes séquentiels classiques.

Cet article est organisé de la manière suivante: la section 2 présente un rapide survol des techniques de modélisation de la parole. La section 3 décrit notre approche originale basée sur un réseau de neurones multicadence. La section 4 expose les résultats expérimentaux de notre système, ainsi que ceux obtenus par des approches plus classiques. La dernière section conclut cet article.

2. Les techniques de modélisation

La majorité des systèmes actuels sont basés sur l'utilisation des modèles de Markov cachés (HMM, *Hidden Markov Model*). Ces modèles statistiques représentent la parole par une séquence d'états. Ceux-ci sont liés par une probabilité de transition et chacun d'entre eux émet un vecteur acoustique avec une certaine distribution de probabilité. Les structures des modèles de Markov cachés sont simplifiées ou limitées par un ensemble d'hypothèses, parmi lesquelles:

- Dans le cas d'un modèle de Markov du premier ordre, le calcul de la probabilité est localisé et rend la modélisation de corrélations à long terme difficile.
- La théorie des modèles de Markov cachés est construite sur l'hypothèse de l'indépendance statistique de chaque trame de parole.
- Il est courant de simplifier les matrices de covariances des modèles en ne gardant que les composants de la diagonale. Cette simplification est justifiée par la difficulté de réestimer les autres composants. Cependant, elle entraîne la limitation suivante: les coefficients des vecteurs acoustiques sont considérés comme statistiquement indépendants.

Une alternative dans les techniques de modélisation est constituée par la méthode connexionniste, plus couramment appelée: les réseaux de neurones artificiels. Dans le cas qui nous préoccupe ici, à savoir la modélisation des signaux de parole, nous parlerons d'un réseau de neurones spatio-temporel. Cette différenciation souligne le fait que, pour une modélisation de signaux dynamiques comme la parole, il est nécessaire d'introduire de la mémoire dans le réseau. Dans ce cadre, plusieurs architectures ont été proposées, en particulier les réseaux (TDNN, *Time Delay Neural Networks*) et les réseaux récurrents (RNN, *Recurrent Neural Networks*). Ces structures permettent de modéliser le rapport entre trames successives par l'adjonction d'éléments de mémoire dans leur réseau. Certaines des limitations HMM sont ainsi contournées. Néanmoins, la modélisation séquentielle reste le point faible de ces architectures connexionnistes.

Une évolution naturelle consiste à construire des modèles hybrides NN/HMM [BM94],[Ben95]. Une

propriété intéressante de cette approche est une modélisation par fonctions discriminantes des classes. En d'autres termes, ce sont les frontières entre les classes qui sont modélisées. Une des conséquences de cette modélisation discriminante est une réduction significative du nombre de paramètres nécessaires. En comparaison, les modèles HMMs, entraînés par le critère du maximum de vraisemblance, procèdent de manière indirecte en modélisant les fonctions de densité de probabilité (les vraisemblances) de manière individuelle à chaque état Markovien.

3. Les réseaux de neurones multicadences

Dans cette section, les réseaux de neurones multicadences (MNN, *Multirate Neural Networks*) sont décrits comme une nouvelle famille de systèmes connexionnistes. La partie originale réside dans l'utilisation explicite des opérateurs de décimation et d'interpolation nécessaires au changement de fréquence d'échantillonnage. L'addition de ces opérateurs permet d'intégrer plusieurs fréquences de travail dans un même réseau. Le résultat de ces contraintes spatio-temporelles est une modularisation implicite du réseau dans un sous-ensemble de représentations localisées. La décimation est utilisée lors de la phase avant, alors que l'interpolation est nécessaire dans la phase de rétro-propagation de l'erreur. L'architecture MNN constitue une extension des réseaux connus sous le nom de *Convolutional Networks* [LB95].

La décimation est définie dans le contexte d'une réduction de la fréquence d'échantillonnage F_x d'un signal $x(n)$ avec un spectre non nul sur la plage de fréquence $|F| \leq F_x/2$. Si la fréquence d'échantillonnage est maintenant réduite d'un facteur D, simplement en prenant chaque Dième échantillon, le résultat sera un signal affecté par le repliement parasite (*aliasing*) des fréquences autour de $F_x/2D$. Pour éviter ce phénomène, l'opération de décimation effectue un filtrage $h_a(n)$ pour réduire la bande spectrale du signal jusqu'à $F_{max} = F_x/2D$ puis, il réduit la fréquence d'échantillonnage d'un facteur D. Les opérateurs de décimation et d'interpolation sont représentés dans la Figure 1.

Les opérateurs de décimation/interpolation et les systèmes multicadences ont été utilisés depuis de nombreuses années dans le domaine du traitement numérique des signaux. Leur intérêt est de réduire le nombre de calculs nécessaires à certaines opérations de filtrage ou, comme dans le cas des ondelettes, d'offrir une meilleure représentation spectro-temporelle.

Dans le domaine non linéaire des réseaux de neurones, les opérateurs de décimation permettent de réduire la complexité et donc de réduire la capacité du réseau. D'autres techniques du même ordre sont l'élagage de certaines connexions (*pruning*) ou le partage de paramètres (*weight-sharing*). L'utilisation de la décimation correspond à l'introduction de connaissances *a priori* et spécifiques au problème traité. L'application de l'architecture MNN se prête directement à la reconnaissance de la parole, si l'on considère la parole comme une information hiérarchique à multi-étages [Gre97]. En effet, le recouvrement temporel des unités linguistiques (phonèmes, groupe de phonèmes et mots) permet d'attacher un étage particulier à chaque entité avec son espace spatio-temporel propre.

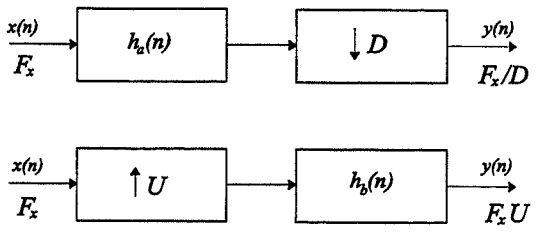


Figure 1: La décimation d'un facteur D et l'interpolation d'un facteur U.

Dans les systèmes de reconnaissance de la parole actuels, l'apprentissage des séquences acoustiques est partiellement limité. Les HMMs, par exemple, définissent la séquence des états de manière a priori: les séquences ne sont pas apprises. Dans le cas des réseaux de neurones, les réseaux perceptrons à multicouches (MLP, *Multi-Layer Perceptrons*) sont directement limités par la taille de leur fenêtre temporelle d'entrée. Les réseaux récurrents sont limités par un phénomène décrit dans [Ben95] de *shrinking gradients* et découvert par Hochreiter (1991) et Bengio, Frasconi et Simard (1993).

L'approche MNN est donc particulièrement intéressante pour modéliser des portions de parole d'une durée plus importante que la trame, le facteur de décimation produisant une multiplication de la durée temporelle.

4. Le système expérimental

Cette partie expérimentale a pour objectif de mesurer la robustesse du modèle MNN par rapport à des systèmes classiques. Cette mesure est définie sur un vocabulaire composé de chaînes de mots. La méthode consiste à entraîner l'ensemble des systèmes sur une première base de données, puis de les tester sur une deuxième possédant des caractéristiques acoustiques sensiblement différentes de la première. La robustesse recherchée est un comportement caractérisé par une dégradation moindre des performances lorsque des différences apparaissent entre la base de données d'entraînement et les conditions de tests.

Les deux bases de données choisies sont de type téléphonique: Swiss-French Polyphone et Computer95. Swiss-French Polyphone a été utilisée pour l'entraînement des systèmes. Le vocabulaire est composé de l'ensemble des chiffres {0 ... 9} et des trois mots additionnels {étoile, dièse, astérisque}. Chaque séquence de 6 mots est prononcée par un locuteur différent. Un premier partage des données est réalisé entre le sous-ensemble d'entraînement de 3302 séquences et le sous-ensemble de test avec 662 séquences. Ceci permet de comparer les systèmes

lorsque les conditions d'entraînement et de test sont identiques. La base de données Computer95, enregistrée dans le cadre d'une exposition, est représentative d'un environnement acoustique bruité. Le vocabulaire est composé uniquement de chiffres avec des séquences d'une longueur de 10 mots. Le nombre total de séquences est de 312 avec 79 locuteurs différents, donc approximativement 4 séquences par locuteur.

Les tailles respectives des différentes bases de données déterminent si les différences mesurées entre les taux de reconnaissance sont statistiquement significatives.

4.1. La description de l'architecture

Avant de décrire l'architecture générale de la Figure 3, il est nécessaire de définir les blocs des fonctions internes.

- Le bloc STNN, *SpatioTemporal Neural Network*, du réseau de neurones spatio-temporel est constitué d'un MLP avec une ligne à retards comme entrée. La fonction d'activation est une sigmoïde avec l'entropie relative comme fonction de coût [BM94].
- Le bloc SMNN, *Self-Modularization Neural Network*, de la Figure 2 est composé d'un bloc STNN suivi de l'opérateur de décimation. Chacun bloc SMNN représente un étage de l'architecture.

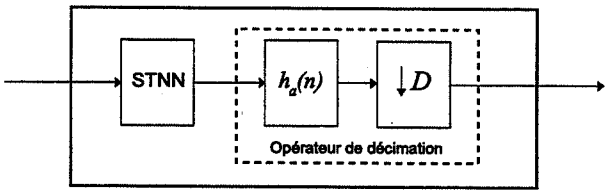


Figure 2: Le bloc SMNN.

La Figure 3 donne une vue schématique d'un système de reconnaissance de mots. Le système est composé de trois étages. Celui du milieu contient un bloc SMNN complet avec un facteur de décimation de quatre. Dans cette première réalisation[1], la rétro-propagation a entraîné chaque étage séparément. Cette procédure d'entraînement devrait être affinée par une rétro-propagation globale sur l'ensemble du réseau. Cette dernière étape n'a pas été réalisée dans le cadre de l'expérimentation.

Pour chaque trame, un vecteur de paramètres RASTA [HM94] est extrait avec leurs premières dérivées ainsi que la première dérivée et deuxième dérivée du logarithme de l'énergie, totalisant 26 coefficients en tout. Ces vecteurs sont ensuite normalisés avec une moyenne nulle et un écart-type égal à l'unité. Sans rentrer dans les détails du décodage, il faut noter que, le système MNN testé n'utilise pas d'algorithme de programmation dynamique de type DTW ou Viterbi: les mots sont simplement détectés et la séquence des meilleurs candidats est extraite.

[1]Nous parlons de première réalisation pour indiquer que le choix de l'architecture décrite est une sélection *a priori* dans un espace de possibilités.

Table 1: Les spécifications MNN.

Etages	*Entrée*	*Milieu*	*Sortie*
Architecture			
Unités d'entrée	234	468	504
Unités cachées	400	200	100
Unités de sortie	26	28	14
Fenêtre-trames	9	18	18
Décimation	1	4	1
Cibles	phonèmes	groupes	mots
Entraînement			
Trames	1'362'730	1'354'630	385'574
PTC-entraîn.	86.8%	91.4%	96.0%
PTC-validation	85.1%	89.2%	94.4%
Epochs	8	6	3

4.2. La procédure d'entraînement

Chacun des étages (les blocs SMNN) nécessite une assignation de cibles pour l'entraînement supervisé des réseaux MLP. L'opération de segmentation a été effectuée par l'entraînement d'un système HMM accessoire. L'alignement des états Markoviens a été extrait par l'algorithme de Viterbi et les segments ont ensuite été regroupés manuellement pour former les cibles de phonèmes, groupes de phonèmes et de mots. Les cibles doivent encore être alignées dans le domaine temporel spécifique à chaque étage et par conséquent, elles sont également traitées par la décimation.

L'entraînement des réseaux MLP est quelque peu spécifique à la reconnaissance de la parole. En effet, la grande taille des bases de données et leur redondance ont conduit les chercheurs à adopter des procédures spécifiques au domaine [BM94]. Un exemple est résumé ici.

- L'entraînement est basé sur une assignation d'une cible à chaque trame (10ms). Chaque fenêtre temporelle de l'entrée du réseau est associée à une cible. Les performances sont mesurées en PTC, le pourcentage de trames correctement reconnues.
- La base d'entraînement est partagée en deux parties: le sous-ensemble d'entraînement avec 2700 séquences et le sous-ensemble de la validation croisée avec 602 séquences, chaque séquence étant prononcée par un locuteur différent.
- Seulement quelques passes (*epochs*) sont nécessaires pour que l'algorithme de rétro-propagation stochastique *on-line learning* converge vers un minimum local.
- A chaque passe, une simple stratégie adaptative décroît le taux d'apprentissage. Lorsque les performances stagnent sur le sous-ensemble de la validation croisée, le coefficient est divisé par deux.

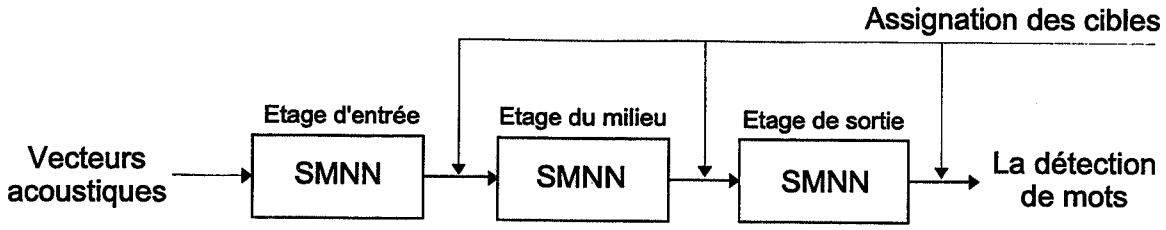


Figure 3: Une architecture MNN à trois étages.

4.3. Les résultats expérimentaux

Tous les systèmes ont été entraînés sur les données de Swiss-French Polyphone. Aucun des systèmes n'a eu accès à la partie de test de Swiss-French Polyphone ni à la base de données Computer95. Les systèmes ont été testés sur les deux bases de données suivantes: (i) l'ensemble de test de Swiss-French Polyphone qui caractérise des conditions acoustiques identiques entre l'entraînement et le test, (ii) la base Computer95 qui détermine un exemple type de différences marquées entre les conditions d'entraînement et de test.

Les systèmes testés sont les suivants:

- V1, le système de la firme VCS est considéré comme l'un des meilleures dans la catégorie de la reconnaissance des séquences de chiffres.
- V2 est un système proposé par la firme DASA.
- ALGN est le système utilisé pour créer l'alignement des modèles MNN. Les résultats de ce système sont donnés à titre indicatif. En effet, la simplicité des modèles (une gaussienne par état) ne le rend pas représentatif des modèles de Markov cachés.
- REF est un système à multi-gaussiennes avec un modèle HMM par mot. Lors de l'entraînement, la complexité des modèles a été augmentée progressivement avec un incrément de deux jusqu'à 11 gaussiennes, à partir de cette complexité de modèles, l'augmentation des performances stagne sur le sous-ensemble de la validation croisée. Ce système utilise les mêmes vecteurs acoustiques RASTA que le système MNN. Il possède également la même résolution acoustique.
- Le dernier système est celui du modèle MNN à multi-étages avec les spécifications définies dans le Tableau 1.

Table 2: Les taux de reconnaissance des chaînes pour les systèmes testés.

	Swiss Polyphone	*Computer95*
V1	79.3%	30.5%
V2	63.8%	9.9%
ALGN	76.4%	12.5%
REF	83.1%	19.9%
MNN	81.3%	49.0%

Les résultats du Tableau 2 permettent de tirer les conclusions suivantes: bien que les performances des trois meilleurs systèmes soient très proches lors de conditions identiques entre l'entraînement et le test, ils manifestent par contre un comportement significativement différent lorsque les conditions acoustiques de test diffèrent de celles de l'entraînement. La robustesse du système MNN est ainsi clairement mise en évidence.

5. Conclusion

L'approche d'une modélisation intrinsèquement robuste ne nécessite pas d'hypothèse a priori sur le type de perturbations, contrairement aux systèmes adaptatifs. Notre système MNN, basé sur des modèles hiérarchiques à multi-étages, présente des caractéristiques intéressantes de robustesse à des perturbations qui ne sont pas représentées dans la base de données d'entraînement. En outre, le système décrit est le premier de sa génération, une procédure d'optimisation basée sur un algorithme génétique distribué est actuellement mise en place pour explorer l'espace des possibilités offertes par son architecture.

Bibliographie

[Ben95] Y. Bengio. *Neural Networks for Speech and Sequence Recognition.* International Thomson Computer Press, Berkshire House, High Holborn, London WC1V7AA, UK, 1995.

[BM94] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, 3300 AH Dordrecht, The Netherlands, 1994.

[Gre97] Steven Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-à-Mousson, France, April 1997.

[HM94] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Trans. on Speech and Audio Proc.*, 2(4):578–589, October 1994.

[LB95] Y. LeCun and Y. Bengio. Convolutional Networks for Images, Speech, and Time Series. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. The MIT Press, Cambridge, Massachusetts 02142, 1995.

Vers une meilleure modélisation du langage: la prise en compte des séquences dans les modèles statistiques

I. Zitouni, K. Smaïli

LORIA/INRIA-Lorraine
B.P.239 54506 Nancy, France
E-mail: {zitouni, smaili}@loria.fr

ABSTRACT

In natural language, several sequences of words are very frequent. Conventional language models do not adequately take into account such sequences, because they underestimate their probabilities. A better approach consists in modeling word sequences as if they were individual dictionary elements.
In this paper, we present an original method for automatically determining the most frequent phrases. This method is based on information theoretic criteria which insure a high statistical consistency, and on French grammatical classes which include additional linguistic dependencies. We propose also several language models based on theses word sequences. Experimental tests on a vocabulary of *20000* words show that the perplexity is reduced by more than *25%* compared to conventional models. The introduction of these word sequences in our dictation machine MAUD reduces the word error rate by more than *22%*.

1. INTRODUCTION

Une phrase est soumise à différentes contraintes : lexicales liées à la limitation du vocabulaire, syntaxiques et sémantiques régissant l'ordre des mots. Dès lors, toutes les combinaisons possibles de mots ne sont pas observables, loin s'en faut. Pour des raisons syntaxiques et/ou sémantiques, certaines suites de mots forment un groupe homogène. Ces séquences, naturellement de longueur variable, véhiculent souvent soit une idée soit une structure langagière particulière. Afin d'introduire cette notion importante dans les systèmes de reconnaissance automatique de la parole (RAP), nous avons généralisé plusieurs modèles de langage classiques en y incluant cette notion de séquences. Tout d'abord, nous avons considéré l'ensemble de ces séquences comme des unités de la langue. Ensuite, nous les avons ajoutées au vocabulaire de base, construisant ainsi un nouveau vocabulaire. Par conséquent, lors de la prédiction et même de la sélection, les modèles de langage utilisant ce nouveau vocabulaire se fondent sur un historique d'unités où chacune d'entre elles peut être, soit un mot, soit une séquence. Ceci donne aux modèles la possibilité d'utiliser un historique plus important et de mieux prendre en compte le rôle prédictif de ces séquences. Les modèles de langage correspondants sont également capables de prédire la totalité d'une séquence, et de ne plus se limiter à la prédiction d'un seul mot.
L'introduction de ces séquences peut augmenter le nombre d'unités (mots ou séquences de mots) peu fréquentes, ce qui risque de réduire les performances des modèles de langage. Les séquences ne doivent donc pas être introduites arbitrairement dans le vocabulaire initial. Nous présentons dans ce papier une nouvelle approche d'extraction de séquences de mots se fondant sur la perplexité et l'information mutuelle. Pour mieux prendre en compte les contraintes syntaxiques de la langue, nous utilisons également un ensemble de classes syntaxiques. A la différence des approches d'extraction des séquences, proposées par R-L. Mercer [1], E. Giachin [2], B.Suhm [3] et S. Deligne [4], celle-ci permet l'utilisation de grands vocabulaires sans nécessiter d'énorme capacité de calcul. Les résultats d'évaluation, présentés dans ce papier, sont estimés sur un vocabulaire de *20000* mots. Nous présentons les performances apportées par l'utilisation de ces séquences dans les nouveaux modèles de langage que nous proposons : n-SeqGrammes et n-SeqClasses correspondant respectivement à l'extension des n-grammes, et des n-classes. Pour mieux appréhender les contraintes langagières qui existent entre le mot courant et un historique lointain, nous proposons un modèle fondé sur la notion de triggers de séquences. Si une séquence A est fortement corrélée avec une séquence B, le couple (A-B) est considéré comme un trigger de séquence. Ainsi, si A apparaît dans l'historique, la vraisemblance de B est renforcé. L'originalité de cette approche, par rapport à celle couramment utilisée dans la littérature [5], est qu'elle se serve des séquences ; ce qui est linguistiquement plus riche. En effet, si on demande à une personne de prédire la suite de la phrase tronquée "le nom du président actuel est", il est plus naturel pour elle de répondre : "Jacques Chirac" plutôt que "Jacques".

2. INTERET DES SEQUENCES EN RAP

L'utilisation des séquences de mots, comme des unités à part entière dans un vocabulaire, permet aux modèles de langage de tenir compte de contraintes langagières supplémentaires ; ce qui accroîtra leurs performances. En effet, l'utilisation de ces séquences comporte de nombreux avantages :
L'historique traité par les modèles de langage à contexte fixe (n-grammes, n-classes, etc.) devient variable et plus long.
La prononciation orale d'une expression contenant plusieurs mots est souvent différente de celle où l'on prononce un à un (d'une façon indépendante) les mots qui la composent. Ainsi, le modèle à base de séquences apporte un plus au niveau acoustique par le biais de l'utilisation d'un modèle de prononciation spécifique à chaque séquence de mots.
Les résultats fournis par un système de RAP permettent

d'obtenir des phrases plus cohérentes au niveau linguistique que celles obtenues par des modèles à base de mots. Ceci est dû au fait que les séquences construites automatiquement ont souvent une structure linguistique viable.

3. ALGORITHME D'EXTRACTION DE SEQUENCES DE MOTS

L'algorithme d'extraction de séquences que nous proposons est entièrement automatique. Il se fonde sur un jeu de classes qui peut être déterminé soit manuellement soit automatiquement. Dans notre cas, nous utilisons un ensemble de classes syntaxiques V_c du français, construites manuellement par des experts ; un mot peut appartenir à plusieurs classes. L'algorithme commence ainsi par étiqueter automatiquement le corpus de mots *W*, avec l'ensemble des classes syntaxiques, construisant ainsi le corpus de classes *C*. Etiqueter le corpus revient à résoudre syntaxiquement le texte, autrement dit affecter à chaque mot sa classe syntaxique contextuelle. Nous utilisons pour ceci un algorithme de type Viterbi [6]. Ensuite, l'algorithme identifie dans *C* tous les couples de classes ou de séquences de classes adjacentes réduisant la perplexité. A partir de ces séquences de classes, on extrait les séquences de mots correspondant dans *W*. Ces unités ainsi construites constituent le résultat final de cet algorithme (les séquences de mots recherchées). Nous ne prenons en compte que les séquences de classes dont les mots correspondant appartiennent au vocabulaire *V*. Pour contrôler la convergence de l'algorithme, le nombre maximum de mots (q) dans une séquence est fixé *a priori*. Afin de limiter le nombre de séquences, l'algorithme prend en compte seulement les séquences de classes qui apparaissent fréquemment et qui ont une information mutuelle supérieure à un certain seuil T_j :

$$T_j = p \max_{c_i \in C, c_j \in C} J(c_i, c_j) \quad (1)$$

où p est un coefficient prédéfini et $J(c_i,c_j)$ représente l'information mutuelle du couple de classes ou de séquences de classes en argument :

$$J(c_i, c_j) = \log \frac{N(c_i, c_j) T}{N(c_i) N(c_j)} \quad (2)$$

où $N(.)$ désigne l'occurrence de l'argument et T la taille du corpus. Une information mutuelle assez grande entre deux composants c_i et c_j indique que ces séquences ne sont pas apparues l'une à coté de l'autre, par un pur hasard, mais qu'elles constituent soit un groupe soit une partie d'un groupe au sens linguistique.

La prise en compte ou non d'une séquence de classes et d'une séquence de mots dépend du nombre minimal d'occurrences de chacune d'entre elles. Ces deux valeurs minimales sont notées T_{min} et T_{occ} dans l'algorithme de construction de séquences donné ci-dessous :

1) Déterminer, à partir du corpus C, les couples de classes ou de séquences de classes $c_i c_j$ ayant une information mutuelle supérieur à T_j et un nombre de constituants inférieur à q.

2) Ajouter ces séquences candidates de classes à V_c, construisant ainsi V_c'. Remplacer les par de nouvelles étiquettes dans *C*, formant ainsi *C'*.

3) A partir de ces séquences de classes, extraire les séquences de mots correspondantes dans le corpus *W*.

4) Ajouter les séquences $\{s_i s_j\}$, obtenues dans la troisième étape, de fréquence supérieure à T_{occ} dans le vocabulaire *V*, construisant ainsi *V'*. Mettre à jour *W*, en replaçant ces séquences par de nouvelles étiquettes, formant ainsi *W'*.

5) Utiliser les vocabulaires *V'* et V_c' pour calculer la perplexité du modèle en utilisant le nouveau vocabulaire *V'*.

6) Si la perplexité diminue : affecter *V'* à *V*, *W'* à *W*, *C'* à *C* et retourner à la première étape.

7) Sinon, extraire des séquences ajoutées à *V'* celles qui réduisent la perplexité, une fois considéré comme unité.

Pour extraire les séquences de la septième étape, nous trions ces séquences en fonction de la valeur de l'information mutuelle des classes correspondantes et nous procédons par dichotomie sur chacune des deux moitiés. En effet, si la moitié en cours réduit la perplexité, nous l'additionnons à *V*. Sinon, nous la divisons en deux sous-ensembles, que nous traitons de nouveau.

La valeur de la perplexité est évaluée à chaque itération avant et après l'ajout des séquences candidates. Sur un corpus de T mots, cette valeur est calculée par un modèle biclasses comme suit :

$$PP = 2^{\frac{1}{T}\log_2 P(w_1) \prod_{i=2}^{T} P(w_i / w_{i-1})}. \quad (3)$$

où $P(w_i/w_{i-1})$ est définie par la formule suivante [7] :

$$P(w_i / w_{i-1}) = \sum_{c_i \in C_{w_i}} p(w_i / c_i) \times \sum_{c_{i-1} \in C_{w_{i-1}}} p(c_{i-1} / w_{i-1}) p(c_i / c_{i-1}) \quad (4)$$

où C_{w_i} est l'ensemble des classes syntaxiques auquel le mot w_i peut appartenir.

Pour estimer la valeur de la perplexité, il n'est pas nécessaire de parcourir, à chaque itération, toutes les unités (mots ou séquences de mots) du corpus. Il suffit de réestimer la vraisemblance des nouvelles séquences, de leurs unités voisines et d'extraire ainsi la nouvelle valeur de la perplexité [7]. Ce traitement améliore considérablement le temps de calcul nécessaire pour le déroulement de cet algorithme. Lorsque nous remplaçons les séquences de mots candidates par des unités, la taille du corpus se réduit et celle du vocabulaire s'agrandit. Pour pouvoir comparer les valeurs de perplexité, avant et après l'ajout des séquences candidates, nous gardons constant la valeur de T lors du calcul de la perplexité [8].

Pour extraire des séquences longues (exemple, *qu'elle heure est t'il*), plusieurs séquences courtes vont être générées (exemple, *qu'elle heure*). Certaines de ces séquences courtes ne sont plus utiles. Si le remplacement d'une séquence courte par ses constituant réduit la perplexité, celle-ci est supprimée.

Nous présentons dans la figure 1, la convergence en terme de perplexité de l'algorithme cité ci-dessus, en fonction du nombre d'itérations et de la valeur de q (le nombre maximum de mots dans une séquence). Les résultats montrent que cette procédure atteint son optimum pour une valeur de q égale à *6*, avec seulement *10* itérations.

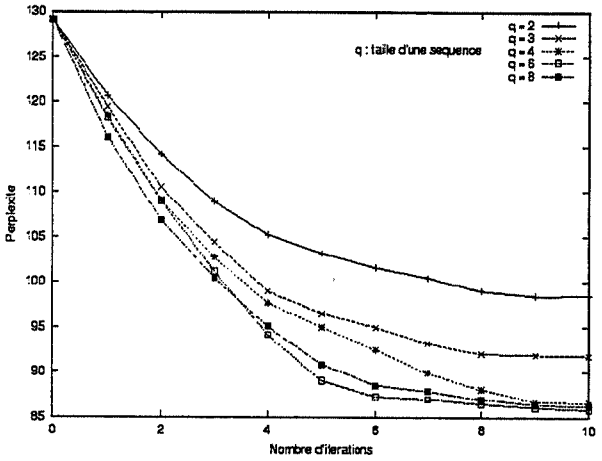


Figure 1 : convergence en terme de perplexité de l'algorithme d'extraction de séquences, à chaque itération, avec différentes valeurs de q (nombre maximum de mots dans une séquence).

4. MODELES A BASE DE SEQUENCES

Les modèles de langages couramment utilisés en reconnaissance automatique de la parole sont les n-grammes et les n-classes. En utilisant ces modèles, la vraisemblance d'un mot ne dépend que des *n-1* derniers mots de l'historique. Pour évaluer les performances des séquences, nous avons construit de nouveaux modèles n-grammes et n-classes utilisant le vocabulaire, résultat de l'algorithme présenté ci-dessus. Ce vocabulaire contient, en plus des mots du lexique de base, l'ensemble des séquences clés de mots ; ces séquences sont considérées comme des unités du vocabulaire. Nous notons par n-SeqGrammes et n-SeqClasses la généralisation des modèles n-grammes et n-classes respectivement, en y introduisant les séquences.

Pour prendre en compte les contraintes linguistiques existantes dans un historique distant, nous avons utilisé les modèles cache et triggers où l'unité de traitement est, soit un mot, soit une séquence de mots. Soit un trigger (M_1-M_2) ; dans la littérature, les unités M_1 et M_2 se limitent aux mots. Dans le modèle que nous proposons, elles peuvent être soit des mots soit des séquences de mots. Nous nommons SeqTriggers et SeqCache respectivement les modèles triggers et cache utilisant des séquences.

L'information mutuelle est une bonne mesure, permettant d'évaluer la quantité d'information apportée par un mot M_1 pour la prédiction d'un autre mot M_2. Nous avons ainsi utilisé cette mesure pour extraire les K meilleurs triggers (M_1-M_2) qui réduisent la perplexité [5]. Cette perplexité est calculée pour un modèle de langage, résultat d'une interpolation linéaire entre les modèles n-grammes, cache et triggers :

$$P(s/h)=\alpha P_{gram}(s/h)+\beta P_C(s/h)+\delta P_T(s/h) \quad (5)$$

où α, β et δ représentent les paramètres d'interpolation .

5. EVALUATION

5.1 Description des données

Le corpus utilisé (LeM) pour l'apprentissage des modèles de langage contient *43* millions de mots, représentant deux années (*87-88*) du journal « Le Monde ». Pour estimer les modèles n-classes et n-SeqClasses, nous avons utilisé un ensemble de *233* classes syntaxiques. Ces classes sont construites manuellement par des experts [6]. Le corpus de classes est celui obtenu par l'étiquetage de LeM. Le vocabulaire de base contient les *20000* mots les plus fréquents dans LeM.

5.2 Evaluation en terme de perplexité

Nous avons évalué nos modèles sur un corpus de test, contenant *5* million de mots qui n'ont pas servi à l'apprentissage. Signalons que l'application de l'algorithme d'extraction automatique de séquences nous a permis d'inclure 4000 nouvelles séquences dans le dictionnaire. Nous avons réparti nos modèles en deux catégories : la première (C_1) fondée sur l'utilisation exclusive des mots et la seconde (C_2) sur l'utilisation de séquences de longueur variable. La catégorie C_1 contient les modèles : bigrammes (P1), trigrammes (P2), biclasses (P3), triclasses (P4), un modèle obtenu par une interpolation linéaire entre bigrammes, cache et triggers (P5) et un modèle résultat d'une interpolation linéaire entre trigrammes, cache et triggers. La catégorie C_2 comporte les 2-SeqGrammes (PS1), 3-SeqGrammes (PS2), 2-SeqClasses (PS3), 3-SeqClasses (PS4), un modèle obtenu par une interpolation linéaire entre 2-SeqGrammes, SeqCache et un modèle résultat d'une interpolation linéaire entre 3-SeqGrammes, SeqCache et SeqTriggers. Pour apprendre l'ensemble de ces modèles, que se soit pour C_1 ou C_2, nous utilisons la méthode de Katz [9]. Nous exposons dans la table 1, les valeurs de perplexité de ces modèles de langage.

C_1	P1	P2	P3	P4	P5	P6
PP	*121.53*	*74.65*	*129.14*	*81.94*	*117.53*	*72.69*
C_2	PS1	PS2	PS3	PS4	PS5	PS6
PP	*83.63*	*63.96*	*85.87*	*72.12*	*80.00*	*60.95*

Table 1 : Comparaison en terme de perplexité des modèles de langage avec et sans séquences.

Une comparaison, en terme de perplexité entre l'ensemble de ces valeurs, montre que l'introduction des séquences améliore d'environ *25%* les performances des modèles de langage classiques.

5.3 Le système MAUD

L'amélioration des performances d'un modèle de langage en terme de perplexité ne suffit pas pour affirmer la supériorité d'un modèle par rapport à un autre, il faut

absolument le tester dans un cadre de reconnaissance. Pour ce faire ,nous avons intégré les modèles de langage que nous avons développé dans notre machine à dicter MAUD [10]. La version de base de MAUD se fonde sur un modèle probabiliste de langage et procède en *3* étapes : identification du genre du locuteur, construction d'un treillis de mots à partir d'un modèle acoustique dépendant du sexe et extraction de la meilleure hypothèse. Les deux premières passes utilisent un algorithme de type Viterbi-Bloc avec un modèle de langage bigrammes. La troisième étape se fonde sur un algorithme de recherche en faisceaux, sur un modèle trigramme et sur le score acoustique des mots reconnus lors de l'étape précédente.

Chaque phonème est représenté par un modèle de Markov de second ordre à trois états (HMM2) [11]. Un mot est ainsi défini par la concaténation des HMM2 des phonèmes qui le composent. Dans le cas d'une séquence, nous introduisons un HMM2 de silence optionnel entre les mots qui la composent. Pour estimer les HMM2 de phonèmes, nous utilisons le corpus de parole Bref80 [12].

Plusieurs versions du système ont été développées : **M1** représentant la version de base exposée ci-dessus ; **MS1** utilise un modèle 2-SeqGrammes lors de la deuxième passe et un modèle 3-SeqGrammes lors de la troisième passe ; **M2** est identique à M1 avec la différence qu'elle utilise un modèle biclasses et triclasses au lieu des modèles bigrammes et trigrammes respectivement ; **MS2** dans laquelle nous introduisons les séquences et nous remplaçons les modèles biclasses et triclasses de M2 par des modèles 2-SeqClasses et 3-SeqClasses respectivement ; **M3** ressemble à la version M1 et utilise un modèle résultat d'une interpolation linéaire entre les modèles trigrammes, cache et triggers ; **MS3** est identique à MS1 avec la différence qu'elle utilise en plus les modèles SeqTriggers et SeqCache lors de la troisième passe.

Nous présentons dans la table 2, le taux d'erreur de chacune des versions de MAUD citée ci-dessus. Nous estimons ce taux sur les 300 phrases de tests, fournies lors de la campagne de l'AUPELF-UREF d'évaluation des systèmes de reconnaissance automatique de la parole.

	M1	MS1	M2	MS2	M3	MS3
WER	*38%*	*29.9%*	*43%*	*33.5%*	*36.6%*	*28.1%*

Table 2 : Taux d'erreur (WER) des différentes versions du système MAUD avec et sans séquences de mots.

Les résultats de reconnaissance montrent que l'utilisation des séquences a permis de faire baisser d'environ *22%* le taux d'erreur de MAUD.

6. Conclusion et perspectives

Une nouvelle approche a été proposée dans ce travail, permettant d'améliorer les performances des modèles de langage et de celles de notre système de reconnaissance automatique de la parole. En utilisant cette approche, les modèles de langage sont évalués en utilisant un lexique de séquences de mots. Ces séquences sont extraites automatiquement, en se basant sur des critères d'optimalité connus en théorie de l'information : la perplexité et l'information mutuelle. Comparativement au modèle n'utilisant que les mots, le modèle à base de séquences a permis de réduire la valeur de la perplexité de *25%* et d'améliorer le taux de reconnaissance de MAUD d'environ *22%*. La majorité de nos séquences ont une structure syntaxique correcte. De plus, beaucoup d'entre elles sont de nature sémantique. L'originalité et la faisabilité de cette méthode sont dues au fait que le système générateur de séquences de mots est fondé sur des séquences de classes construites automatiquement. Ces séquences doivent elles-mêmes vérifier un certain nombre de contraintes (une information mutuelle supérieure à un certain seuil, un nombre limité de constituants et un contrôle permanent des séquences par le biais de la perplexité) avant d'être candidates à générer des séquences de mots.

Pour améliorer les performances de cette approche, nous comptons l'utiliser avec le modèle multigrammes [4]. Nous pensons également appliquer cette approche dans d'autres domaines : construction des classes d'équivalence sémantiques, identification thématique, traduction automatique, ...

References

[1] Jelinek F. "Self-Organized Language Modeling for Speech Recognition". In Readings in Speech Recognition, pp. 450-506. Ed. A. Waibel and K. F. Lee. Morgan Kaufmann, 1989.

[2] Giachin E. "Phrase Bigrams for Continuous Speech Recognition". ICASSP95, Detroit, pp. 225-228, 1995.

[3] Suhm B. and Waibel *A.* "Towards Better Language Models for Spontaneous Speech". ICSLP94, Yokohama, pp. 831-834, 1994.

[4] Deligne S. and Bimbot F. "Language Modeling by Variable Length Sequences: Theoretical Formulation and Évaluation of Multigrams". ICASSP95, Detroit, pp. 169-172, 1995.

[5] Tillmann C., Ney H., "Selection Criteria for Word Trigger Pairs in Language Modeling". In Grammatical Inference: Learning Syntax from Sentences, Springer, 1996.

[6] Smaïli K., Zitouni I., Charpillet F. and Haton J. P. "An Hybrid Language Model for a Continuous Dictation Prototype". Eurospeech97, Rhodes, pp. 2723-2726, 1997.

[7] Zitouni I. "Modlisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaire : application à MAUD". Thèse de l'université Henri Poincarré, Nancy, 2000.

[8] Adda G. et al., "Text Normalisation and Speech Recognition in French". Eurospeech97, Rhodes, 1997.

[9] Katz M. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer". IEEE Trans. ASSP, 35 (3), pp. 400-401, 1987.

[10] Fohr D, Haton J. P., Mari J. F., Smaïli K., Zitouni I. "MAUD: Un prototype de machine à dicter vocale". Actes 1$^{\text{ères}}$ JST FRANCIL, Avignon, pp. 25-30, 1997.

[11] Mari J. F., Haton J. P., Kriouile A. "Automatic Word Recognition Based on Second-Order Hidden Markov Models". IEEE Trans. ASSP, 2(1), pp. 22-25, 1997.

[12] Lamel L., Gauvain J-L., Eskenazi M. "BREF: a Large Vocabulary Spoken Corpus for French". Eurospeech91, Gênes, 1991.

Utilisation de treillis synchrones pour la reconnaissance vocale à partir de références acoustiques uniques

S. Peillon[1], *A. Ferrieux*

France Télécom R&D
2, av. Pierre Marzin – 22300 Lannion, France
Tél.: ++33 (0) 2 96 05 33 06
Mél: {stephane.peillon, alexandre.ferrieux}@francetelecom.fr

ABSTRACT

This paper reports on spoken pattern based speech recognition approach. We develop a solution for interactive voice/voice applications without predefined transcribed vocabulary. The method uses a two-layer decoding scheme, where the intermediary representation of speech is an indexing pass in symbolic units such as phonemes or data-driven phone-like units, which makes the system vocabulary-independent. Phonemes are well known for their discriminative ability between words, while data-driven phone-like units don't need any labelled database for the training phase. Moreover, the use of synchronized lattices at the intermediary level improves the discriminative performance, compared to one-best sequences of phonemes, and makes the comparison algorithm faster than with standard lattices.

1. INTRODUCTION

Des demandes d'applications croissantes dans le domaine des télécommunications amènent de plus en plus l'utilisateur à définir son propre vocabulaire, quelquefois pour une durée très éphémère – entrées de répertoires téléphoniques pour une numérotation automatique, définition de mots clés (noms propres) pour le classement de messages vocaux... Une ergonomie optimale voudrait que l'utilisateur ne prononce qu'une seule fois les mots à reconnaître, et ne soit pas contraint à un quelconque vocabulaire. Dans ce document, nous allons présenter un système capable de comparer des réalisations acoustiques d'un même mot, prononcées par différents locuteurs.

Ce papier s'inscrit dans le cadre de l'indexation vocale de la parole, appliquée à la détection de clés vocales dans des enregistrements téléphoniques. Nous présentons ici un travail préliminaire visant à la reconnaissance de mots à partir de références acoustiques, mais qui pourra facilement être étendu par la suite à une tâche de détection (sans modèle poubelle). Nous ciblons donc une bibliographie adaptée à la recherche de clés dans des enregistrements. Nous citons tout d'abord les travaux de [Vin68] et de [Hig85] qui utilisèrent une technique de comparaison de mots à base d'alignement temporel dynamique sur des références acoustiques. Nous allons utiliser ce même principe, mais à partir d'enregistrements décodés en phonèmes [Pei98] ou unités pseudo-phonétiques [Fon97]. L'alignement dynamique s'effectue alors sur des symboles (meilleure homogénéité des scores) ; la détection d'une clé consiste à rechercher une sous-chaîne symbolique au lieu de la chaîne entière. Pour se rendre indépendant du vocabulaire, on utilisera des symboles phonétiques, capables de composer tous les mots de vocabulaire possibles. [Jun96] [Pei98] [Tan98] présentent de telles techniques de reconnaissance, mais qui sont relativement sensibles aux erreurs de reconnaissance phonétique. [Gel96] [Jam94] eux, utilisent des treillis ou graphes de phonèmes au lieu de séquences phonétiques. Ces treillis fournissent différentes hypothèses phonétiques qui compensent ainsi une part importante des erreurs de reconnaissance. Les performances sont meilleures, mais les algorithmes plus coûteux en temps, ne se généralisent pas simplement à la reconnaissance de mots à partir de références acoustiques puisque les mots à reconnaître (transcription phonétique exacte) doivent être recherchés dans des graphes d'hypothèses. Enfin, nous citerons aussi les systèmes à base de HMM de mots tels que [Ros90], où l'apprentissage d'un modèle propre à chaque mot de vocabulaire ne permet pas la recherche d'un mot nouveau défini par l'utilisateur.

Pour reconnaître des mots à partir de références acoustiques, nous allons au préalable décoder à l'aide d'unités phonétiques aussi bien le mot à reconnaître que les références, de façon à les rendre indépendants du locuteur et du vocabulaire. Ensuite nous pourrons comparer ces représentations symboliques entre-elles pour la reconnaissance d'un mot. Plutôt que d'utiliser des séquences de phonèmes sensibles aux erreurs de reconnaissance [Pei98], nous allons émettre plusieurs hypothèses symboliques par segment. Nous avions préalablement présenté une technique de détection de clés vocales dans des enregistrements vocaux [Fer99] à partir de treillis synchrones de phonèmes. Nous allons dans cet article comparer les résultats de cette méthode sur une tâche de reconnaissance de mots, non plus avec des symboles phonétiques, mais avec des *classes pseudo-phonétiques* générées automatiquement (**CPP**). Ces classes ont pour

[1] convention C.I.F.R.E. avec Alcatel T.I.T.N. Answare, 30, rue Bahon Rault, 35000 Rennes, France

objectif d'être calculées sans supervision à partir d'un quelconque corpus de parole non étiqueté, afin de réaliser un système de reconnaissance s'adaptant à une nouvelle langue avec un coût moindre.

Nous développons un système de reconnaissance à deux étages. Le premier étage est une passe de transformation des mots à comparer sous la forme de treillis synchrones de phonèmes ou de CPP, présentés respectivement dans les sections 2 et 3. La section 4 décrit le deuxième étage : l'algorithme de comparaison des treillis, à base de DTW. Enfin, les sections 5 et 6 présentent l'évaluation des résultats.

2. TREILLIS SYNCHRONES DE PHONÈMES

Le treillis synchrone est un enrichissement du décodage phonétique d'un mot. Au lieu de représenter ce mot sous la forme d'une séquence de phonèmes (décodage), nous le représentons sous la forme d'un treillis où les différentes hypothèses phonétiques sont synchrones à une segmentation du mot, issue d'un décodage préalable. Cela permet de connaître la proximité des autres phonèmes au sein de chaque segment, pour compenser en grande partie les erreurs de substitution de phonèmes. Si dans cette section nous nous concentrons sur les phonèmes, nous présentons dans la section suivante des treillis synchrones de classes issues d'un découpage automatique de l'espace acoustique en classes gaussiennes.

Pour rester indépendant du vocabulaire, les phonèmes semblent parfaitement adéquats puisqu'ils permettent de composer tous les mots. De plus, deux mots sont considérés distincts s'ils diffèrent d'un phonème. Pour créer un treillis synchrone de phonèmes, nous procédons en deux étapes. Nous utilisons tout d'abord un système de décodage phonétique (Viterbi) à l'aide d'un ensemble de modèles contextuels de phonèmes bouclés (i.e. à grammaire nulle) pour obtenir la segmentation du mot à traiter. Ensuite, pour chacun de ces segments, nous générons un vecteur de N probabilités a posteriori (hypothèses) $Pr\,(Pho_i|X^{b\text{-}e})$, où $X^{b\text{-}e}$ est l'observation entre les trames *begin* et *end* du segment. Ces probabilités sont obtenues par renormalisation des vraisemblances des N phonèmes. Pour le français, $N = 37$ phonèmes + modèles de silence.

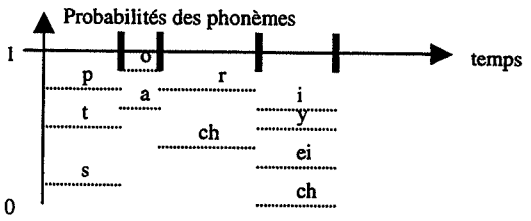


Figure 1: Exemple de treillis synchrone pour le mot *Paris* [p a r i]. (On ne représente que les meilleures hypothèses.)

3. TREILLIS SYNCHRONES DE CLASSES (CPP)

Les CPP sont générées automatiquement par un algorithme de quantification vectorielle sur une base de données non étiquetée.

3.1 Module de création des CPP

Il se base sur un algorithme de minimisation de la distorsion moyenne des classes, de type Lloyd-Max. L'espace acoustique d'entrée est de dimension 18, incluant l'énergie, 8 coefficients cepstraux et leurs dérivées. La création des unités est faite par divisions successives, c'est-à-dire que nous commençons avec une seule classe que nous divisons en deux. Après convergence de ces deux classes, nous redivisons l'une d'entre-elles, et ainsi de suite jusqu'à obtention du nombre de classes désiré. La particularité de notre méthode réside dans une modélisation gaussienne de ces classes, et dans leur critère de division.

L'estimation des gaussiennes est faite à partir des estimateurs classiques de la moyenne et de la variance. Nous choisissons la classe à diviser en fonction d'un critère d'étalement que nous définissons comme le rapport entre la moyenne arithmétique et la moyenne géométrique des vecteurs acoustiques relatifs à la classe. Ce critère, notamment utilisé pour des mesures d'étalement spectral, nous permet de sélectionner la classe la plus étalée vis-à-vis d'une densité gaussienne, privilégiant ainsi le choix d'une classe multimodale à celui d'une classe unimodale.

3.2 Calcul du treillis

Se refusant ici une segmentation phonétique des données de parole, nous procédons au calcul des probabilités des classes trame par trame au lieu de segment par segment. Les probabilités sont les scores des gaussiennes. Par contre, nous aurons recours au calcul d'un score segmental après la phase d'alignement pour compenser la perte procurée par cette analyse synchrone à la trame.

4. ALGORITHME DE COMPARAISON

4.1 Appariement de deux segments acoustiques

Pour le calcul de la dissemblance entre deux mots, nous recherchons l'expression d'une distance entre le treillis du mot à reconnaître et celui de sa référence. Comme détaillé dans la sous-section qui suit, nous effectuons un alignement dynamique entre les deux treillis pour obtenir $Pr\,(réf = mot)$. Pour cela, définissons la probabilité d'aligner deux segments acoustiques représentés par des vecteurs de probabilités, que nous exprimons de la façon suivante :

$$\sum_{i=1}^{N} \Pr\left(Pho_{réf\,i} \mid X_{réf}^{b-e}\right) . \sum_{j=1}^{N} \Pr\left(Pho_i \mid Pho_j\right) . \Pr\left(Pho_{mot\,j} \mid X_{mot}^{b-e}\right)$$

Nous utilisons une matrice de confusion $Pr\,(Pho_i|Pho_j)$ de dimension NxN estimée par apprentissage de façon à minimiser le taux d'erreur de confusion des mots alignés. Elle compense aussi optimalement les erreurs de reconnaissance du générateur d'hypothèses phonétiques, ce qui est particulièrement utile lors d'une reconnaissance indépendante du locuteur sur canal téléphonique.

Notre souhait d'indépendance au vocabulaire nous amène à faire l'hypothèse de non corrélation de deux segments

acoustiques successifs. On pose donc :

$$Pr(réf = mot) = Pr(réf(i(1)) = mot(j(1)) \; ET \; réf(i(2)) = mot(j(2)) \; ET \ldots ET \; réf(i(k)) = mot(j(k)) \; ET \ldots ET \; réf(i(K)) = mot(j(K)))$$

$$= \prod_{k=1}^{K} \Pr\left(réf\left(i(k)\right) = mot\left(j(k)\right)\right)$$

où k est l'indice des segments appariés le long du chemin optimal de l'alignement dynamique, et les fonctions $i(.)$ et $j(.)$ continues croissantes déterminent les indices des segments.

4.2 Alignement dynamique

Notre système de reconnaissance effectue un alignement dynamique sur les segments acoustiques du mot et de la référence avec laquelle on le compare. Ces segments sont dénotés par des vecteurs de probabilités, d'hypothèses phonétiques pour notre première technique, d'hypothèses de classes synchrones à la trame pour la seconde. Contrairement à la comparaison de séquences de phonèmes [Pei98] où nous utilisions une distance d'édition, nous délaissons les coûts d'insertion et d'omission avec les treillis synchrones. En effet, ces derniers, plus riches, précisent des trajectoires phonétiques en émettant des vecteurs d'hypothèses sur chaque segment. Chaque segment représente un tronçon du signal acoustique ; il émet dans le cas d'une sous-segmentation (omission) de fortes hypothèses pour les deux événements phonétiques « fusionnés ». Dans le cas d'une insertion, nous aurons deux vecteurs successifs que nous gardons ; on ne va pas ignorer une partie du signal. En revanche, pour tenir compte des défauts de segmentation, nous allons autoriser les substitutions multiples de segments, c'est-à-dire que l'on va pouvoir apparier un segment sur plusieurs autres.

Pour la DTW sur les segments acoustiques, on utilise donc un graphe de transitions symétrique puisque les mots à aligner sont tous les deux des signaux de parole. Après alignement optimal, nous normalisons le score obtenu par le nombre de segments appariés de façon à définir un seuil de détection/rejet de mots, indépendamment de la référence ou du mot à reconnaître.

Score segmental pour les treillis de classes Dans le cas de classes pseudo-phonétiques qui ne bénéficient pas de modélisation temporelle (contrairement à l'utilisation de HMMs de phonèmes), nous ne voulons pas qu'un segment tel qu'une voyelle assez longue vienne prendre un poids trop important dans le calcul du score, au détriment des autres segments. Nous procédons en deux étapes. La première est l'étape d'alignement ci-dessus, synchrone à la trame. La deuxième, l'étape de calcul du score, va calculer le score final sur le meilleur chemin (backtracking), en fonction de la segmentation de la référence en classes homogènes.

Pour cela, nous faisons l'hypothèse simplificatrice que les trajectoires cepstrales n'oscillent pas constamment d'une classe à l'autre. Nous allons considérer comme appartenant au même segment, toutes les trames consécutives dont les meilleurs hypothèses sont identiques (mêmes CPP). Au sein d'un même segment ainsi identifié, nous normalisons le score cumulé des log-probabilités d'appariement des segments par leur longueur, avant de le multiplier par un facteur de « réduction » R. Cette opération n'est effectuée que pour des segments de plus de R trames. Nous choisissons expérimentalement R=4 (gain de 8 % sur le taux d'égale erreur EER, par rapport à R=∞).

Le score final (somme des scores réduits), est ensuite normalisé par le nombre de segments ainsi retenus. En résumé, nous limitons la taille des segments à R trames. Le résultat est similaire au facteur de compression présenté dans [Fon97].

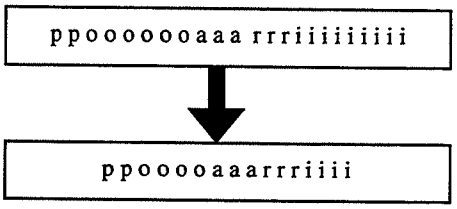


Figure 2: Réduction de segments pour le mot *Paris* [p a r i] dans le cas de treillis de classes, avec R=4. (Nous avons ici remplacé les classes par des phonèmes).

5. CORPUS

Les tests ont été réalisés sur un corpus de courtes phrases (3 à 5 syllabes) enregistrées par téléphone, telles que « Un château hanté » ou « Des nouilles au beurre ». Les enregistrements ont été effectués par 140 locuteurs français prononçant une trentaine de phrases chacun. Au total, 90 phrases distinctes ont donné lieu à 4000 enregistrements équi-répartis entre les locuteurs.

Le corpus d'apprentissage des CPP et des matrices de confusion est parfaitement similaire, mais établi par 140 autres locuteurs et constitué de 90 autres phrases, soit 4000 autres enregistrements.

L'analyse acoustique de ces deux corpus est effectuée toutes les 16 ms, et consiste au calcul de 8 coefficients mel-cepstraux, plus l'énergie et leurs dérivées.

6. RÉSULTATS

Avant toute chose, chaque enregistrement a été décodé sous forme de treillis synchrone de phonèmes ou de classes. Pour le premier cas, les hypothèses phonétiques sont générées à partir d'un ensemble de modèles contextuels de phonèmes (HMMs mono-gaussiens), bouclés pour l'étape de segmentation préalable. Les modèles utilisés sont ceux de la reconnaissance flexible de France Télécom R&D. Pour les CPP, nous avons utilisé le corpus d'apprentissage et construit les classes comme expliqué dans la section 3. Les résultats comparent différents nombres de classes.

Le test en lui-même consiste à aligner des paires de mots

identiques, et des paires de mots différents. Nous tirons du corpus de test environ 80000 paires des deux espèces (soit 160000 paires distinctes), que nous devrons classifier dans la bonne catégorie. Les résultats sont exposés ci-dessous sous la forme de graphique *précision / taux de détection*. Le taux de détection est le nombre de paires de mots identiques correctement classifiées, divisé par le nombre total de paires de mots identiques. La précision est le nombre de paires de mots identiques correctement classifiées, divisé par le nombre total de paires classifiées comme paires de mots identiques.

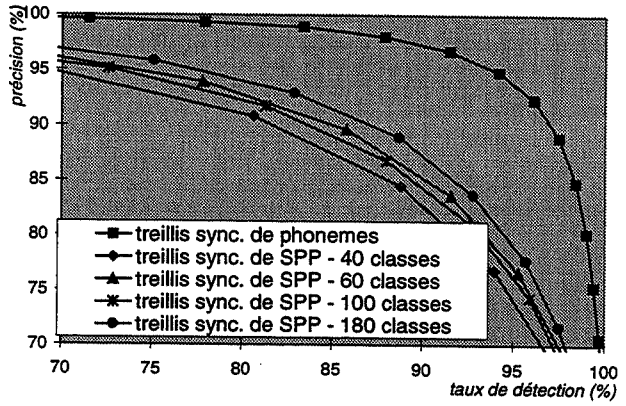


Figure 3: Détection et rejet, respectivement de 80000 paires de mots du corpus allophonique.

Ces courbes sont établies en faisant varier le seuil de détection/rejet, et dénotent les capacités des deux méthodes à mesurer la ressemblance ou la dissemblance de deux mots. L'apport de connaissances humaines et la modélisation temporelle (HMMs) donnent aux treillis d'unités phonétiques des performances supérieures aux treillis de classes.

Un nombre faible de classes donne de faibles performances étant donné le découpage non supervisé du corpus d'apprentissage ; mais plus le nombre de classes augmente, plus nous disposons de classes pour modéliser les événements acoustiques en pseudo-phonèmes. L'objectif est d'identifier la taille d'un dictionnaire perceptuel optimal à partir de CPPs. Il faut cependant veiller à ce qu'un nombre trop élevé de classes (>>180) ne détruise pas la cohésion phonétique de chaque CPP lors de sur-divisions, ne leur permettant alors plus d'absorber autant les sources de variabilité de la parole.

7. Conclusion

Ce travail préliminaire présente une facette de l'indexation vocale pour la reconnaissance de mots à partir de références acoustiques, indépendamment du locuteur et du vocabulaire. L'utilisation de symboles (indexation) à un stade intermédiaire de représentation de la parole (treillis synchrones de phonèmes), permet non seulement de réduire notablement le débit d'informations correspondant à ce niveau de représentation, mais permet aussi d'utiliser des transcriptions phonétiques exactes au lieu d'un décodeur, de façon à pouvoir traiter du texte [Fer99].

Si les treillis synchrones de classes offrent des performances inférieures pour l'instant, des efforts concernant la génération automatique des CPP peuvent être faits, en particulier l'introduction d'une modélisation temporelle, par exemple avec des chaînes de Markov, ou encore en concaténant plusieurs trames successives.

Pour terminer, nous soulignerons la faible complexité algorithmique de notre méthode (de type DTW où la distance locale de deux segments se résume à un produit scalaire après calcul des treillis), ainsi que son faible encombrement mémoire pour le traitement d'un vocabulaire de très grande taille (on n'a besoin que des N modèles phonétiques de la langue, ou des caractéristiques des CPP retenues). Ces travaux semblent donc particulièrement adaptés au développement rapide de toute application nécessitant la reconnaissance ou la détection d'un jeu de mots définis directement par l'utilisateur.

Bibliographie

[Fer99] A. Ferrieux, S. Peillon, " Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval ", ESCA workshop – Accessing information in spoken audio, Cambridge UK, avril 1999

[Fon97] V. Fontaine, H. Bourlard, " Speaker-dependent speech recognition based on phone-like unit models — Application to voice dialing ", I.C.A.S.S.P., 1997

[Gel96] P. Gelin, C. J. Wellekens, " Keyword spotting enhancement for video soundtrack indexing ", I.C.S.L.P., 1996

[Hig85] A. L. Higgins, R. E. Wohlford, " Keyword recognition using template concatenation ", I.C.A.S.S.P., pp. 1233-1236, 1985

[Jam94] D. A. James, S. J. Young, " A fast lattice-based approach to vocabulary independent wordspotting ", I.C.A.S.S.P., pp. 377-380, 1994

[Jun96] J. Junkawitsch, L. Neubauer, H. Höge, G. Ruske, " A new keyword spotting algorithm with pre-calculated thresholds ", I.C.S.L.P., 1996

[Pei98] S. Peillon, A. Ferrieux, " Indexation vocale à vocabulaire illimité à base de décodage phonétique ", J.E.P. Martigny Suisse, juin 1998

[Ros90] R. C. Rose, D. B. Paul, " A Hidden Markov Model based keyword recognition system ", I.C.A.S.S.P., pp. 129-132, 1990

[Tan98] K. Tanaka, H. Kojima, " Speech recognition based on the distance calculation between intermediate phonetic code sequences in symbolic domain ", I.C.S.L.P., 1998

[Vin68] T. K. Vintsjuk, " Recognition of spoken words by methods of dynamic programming ", Kibernetica, vol. 1, pp. 81-88, 1968

Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langage thématiques

Brigitte Bigi, Renato De Mori, Thierry Spriet

Laboratoire d'Informatique d'Avignon
L. I. A. - CERI BP 1228 - 84911 Avignon Cedex 9 - FRANCE
Tél. : ++33 (0)4 90 84 35 36 - Fax : ++33 (0)4 90 84 35 01
Mél : {brigitte.bigi,renato.demori,thierry.spriet}@lia.univ-avignon.fr

Résumé

A robust strategy for dynamic language model selection, based on topic recognition and switching between topic models, is proposed. It is effective because it relies on a small set of well trained topic-dependent language models and on reliable topic recognition. By using perplexity as a performance measure of the LM switching model, a tangible reduction is observed with respect to the use of a single, general, static LM. Different methods are proposed for topic shift detection. Experimental results show that different strategies for topic shift detection have to be used depending on whether high recall or high precision are sought.

Présentation

Dans le cadre d'une amélioration des performances des systèmes de Reconnaissance Automatique de la Parole (RAP), nos travaux visent l'adaptation dynamique de leur composante linguistique. Cette adaptation est réalisée en fonction des thèmes identifiés dynamiquement lors de la dictée. Le principe général de cette approche est illustré par la figure 1. Il consiste à utiliser un modèle généraliste au début de la reconnaissance afin d'initialiser les processus de classification thématique. Par la suite, on adapte le modèle de langage (ML) en fonction du résultat de la classification. Le problème de ce type de modèles est qu'ils nécessitent une grande quantité de *corpus segmenté en thèmes*. Cet article présente un ensemble de solutions possibles qui consistent à segmenter automatique des données qui pourront, par la suite, être utilisées pour l'apprentissage.

En section 1, nous montrons brièvement les méthodes développées pour obtenir une classification thématique sur des textes écrits, puis sur des textes dictés à un système de reconnaissance de la parole. De plus, nous montrerons le potentiel de reconnaissance des modèles de langages thématiques en évaluant le gain de perplexité qu'ils peuvent engendrer. En section 2, nous développons différentes méthodes de segmentation qui génèrent les ruptures thématiques.

1. Classification thématique

La classification thématique est un processus appliqué à un texte et dont le résultat est l'assignation d'un label thématique parmi une liste prédéfinie de labels possibles. Ce travail, déjà présenté en [1], propose l'utilisation de deux ML : un ensemble d'unigrammes thématiques, et un modèle basé sur une mémoire cache [3]. Ce dernier calcule, pour chaque thème, la distance entre une distribution de mots-clés thématiques et le contenu de la mémoire cache. A partir de ces distances, nous obtenons des probabilités associées à chaque thème pour le texte.

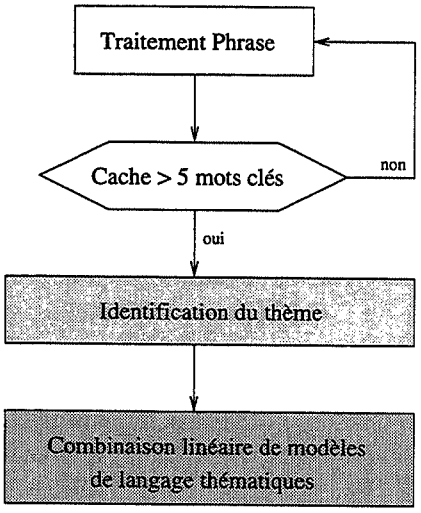


FIG. 1 – Processus d'Adaptation dynamique de modèles de langage thématiques

Les expériences ont été réalisées sur des articles du journal "Le Monde" de 1987 à 1991, avec un vocabulaire de plus de 500 000 entités lexicales. Les sept thèmes retenus sont issus des secteurs rédactionels du journal : *Etranger, Histoire, Sciences, Sports, Economie, Culture, Politique*. Les résultats donnent un taux de classification thématique de plus de 80 % sur un corpus de test composé de 1021 paragraphes.

1.1. Reconnaissance thématique sur corpus dicté

Pour évaluer le taux de classification thématique de textes dictés, nous avons comparé le label assigné par notre système au texte original à celui assigné au texte dicté (sortie du système de reconnaissance). Le corpus de test composé de 97 paragraphes (18000 mots) a été dicté au système *Via Voice 98* d'IBM, un système dépendant du locuteur et à grand vocabulaire. Quatre locuteurs (1 femme, 3 hommes) ont participé à sa constitution. Nous avons obtenus un taux d'erreurs d'environ 35 % sur ce corpus.

Nous obtenons 73 paragraphes correctement étiquetés contre 77 sur les textes de référence correspondants, ce qui signifie que les mots clés sont apparemment bien reconnus. Ce résultat confirme que nos travaux de classification thématique sur l'écrit peuvent être réutilisés dans un système de dictée.

1.2. *Bigramme thématique*

L'expérience consiste à utiliser un modèle de langage thématique sur un corpus de ce thème (ML_t) et d'en comparer la perplexité avec celle d'un modèle de langage général (ML_g). Pour ce faire, on crée un ensemble de bigrammes thématiques et un bigramme général, tous de même vocabulaire (V=10000).

Afin que les modèles thématiques puissent être estimés avec peu de corpus d'apprentissage spécifique au domaine, nous utilisons la combinaison linéaire de chaque bigramme thématique avec le bigramme général :

$$P(w_i|w_j) = \lambda P_g(w_i|w_j) + (1-\lambda)P_t(w_i|w_j)$$

Le coefficient λ, appliqué à ML_g et estimé empiriquement, reflète la qualité du bigramme thématique par rapport à ML_g. Le calcul de la perplexité de ML_g et celui de chaque modèle combiné est réalisé pour chaque thème. Ceux-ci ont été effectués avec le toolkit v2 du CMU ([6]). Les résultats sont reportés en table 1. La deuxième colonne indique la quantité de données utilisées lors de l'apprentissage des modèles thématiques ; le ML_g, quant à lui, a nécessité un autre corpus de 32,3 M de mots. Le thème *histoire* n'ayant pu produire un modèle thématique cohérent, l'estimation du λ a rejeté l'utilisation de ce thème. Un gain de l'ordre de 8,7 % est observé sur les 6 autres thèmes.

TAB. 1 – Comparaisons de perplexités des modèles bigrammes combinés, et du modèle général

Thème	Taille (mots)	λ	PP ML combiné	PP ML général	Gain
Etranger	15,1	0,3	135,9	149	8,8 %
Histoire	0,6	1	-	191,2	-
Science	2	0,6	156,3	174,2	10,3 %
Sport	0,2	0,7	161,5	179,5	10 %
Economie	10,4	0,3	133,9	144,9	7,6 %
Culture	15,8	0,3	160,6	177	9,3 %
Politique	10,1	0,4	138,6	148,4	6,6 %

2. Segmentation thématique

La segmentation thématique a pour but de déterminer automatiquement les frontières thématiques des segments qui composent un document. Les résultats s'expriment sous la forme de taux de rappel et précision, avec :

$$rappel = \frac{Nb\ de\ ruptures\ correctes\ trouvees}{Nb\ de\ ruptures\ a\ trouver}$$

$$precision = \frac{Nb\ de\ ruptures\ correctes\ trouvees}{Nb\ de\ ruptures\ totales\ trouvees}$$

Plusieurs stratégies pour repérer les ruptures de thèmes et pour les sélectionner sont possibles. Une partie de ces travaux est présentée dans [2].

Le corpus de test est composé de 1393 documents générés automatiquement de telle sorte que chacun d'entre-eux est constitué de 3 paragraphes tirés aléatoirement dans nos corpus thématiques. Dans la mesure où la taille des paragraphes qui constituent les documents peut avoir une conséquence importante sur la qualité des résultats, aucune contrainte n'a été appliquée sur celle-ci.

2.1. *Première phase de détection des ruptures candidates*

Deux approches ont été étudiées. La première reprend les travaux développés pour la classification thématique, et utilise un modèle basé sur une mémoire cache. La seconde approche, place les candidats à intervalles réguliers sans tenir compte de la nature du texte.

Le modèle à base de mémoire cache

Comme décrit dans la section précédente, ce modèle est un outil performant pour la classification thématique, c'est pourquoi nous avons voulu l'utiliser dans le cadre de la segmentation. Le contenu de la mémoire cache est une représentation de l'historique puisqu'elle en conserve uniquement les 100 derniers mot clés. La mémoire cache est réinitialisée à chaque nouveau document. Nous utilisons la distance $d_j^*(i)$, distance de Kullback-Liebler normalisée évaluée entre le contenu de la mémoire cache et l'histogramme des mots clés du thème T_j. Pour la classification thématique, le thème assigné au paragraphe est celui dont la distance est la plus petite.

La distance $d_j^*(i)$ est la distance entre la mémoire cache et l'histogramme du thème T_j à la fin de la i-ème phrase du document. Nous nous intéressons à l'évolution de cette distance pour le meilleur thème, au sens de la classification. Cette variation s'exprime par :

$$\delta(i) = d_j^*(i) - d_j^*(i-1)$$

où j est le meilleur thème à la fin de la $(i-1)$ème phrase.

Nous proposons une rupture candidate à chaque variation importante de la distance, c'est à dire quand $\delta(i) > \theta$, où θ est un seuil déterminé expérimentalement.

Les résultats que l'on obtient avec cette méthode sont exprimés dans la table 2. On observe une valeur élevée de rappel, ce qui signifie que peu de frontières thématiques n'ont pas été détectées. Par contre, la faible valeur de précision indique un nombre important de fausses alarmes. Les différentes valeurs de θ ne font que faire varier proportionnellement les valeurs de rappel et de précision.

TAB. 2 – Repérage de ruptures thématiques par le modèle à base de mémoire cache

θ	0,001	0,0014	0,0018	0,002	0,003
Rappel	0,9091	0,8664	0,8152	0,789	0,612
Précision	0,1261	0,1747	0,2309	0,2608	0,4047

Méthode systématique

A fin de comparaison, nous avons aussi utilisé une méthode systématique. Elle place arbitrairement un candidat toutes les N phrases. Cette méthode permet avec $N = 1$ de trouver toutes les ruptures (rappel=1), avec la précision minimale.

Les résultats sont donnés dans la table 3. Comme on pouvait s'y attendre, on constate qu'ils sont moins bons que ceux de la méthode précédente.

TAB. 3 – Repérage des ruptures par le placement d'un candidat toutes les N phrases

N	1	5	10
Rappel	1	0,8391	0,6816
Précision	0,0249	0,1084	0,1805

2.2. Sélectionner les candidats

On définit un segment comme étant la portion de texte comprise entre deux ruptures candidates. Dans la première étape, on a exposé des méthodes pour repérer un ensemble de ruptures candidates qui peuvent représenter la frontière thématique entre deux segments de texte. On observe donc de forts taux de rappel, ce qui satisfait notre objectif. Dans cette deuxième étape, on développe plusieurs méthodes qui devront sélectionner les candidats en minimisant les fausses alarmes, sans trop perdre de la valeur de rappel.

Utilisation de la distance du modèle cache

En utilisant le modèle à base de mémoire cache pour leur classification, il est possible que deux segments successifs soient étiquetés avec le même thème. Les ruptures de thèmes sont alors définies lorsque deux labels thématiques différents sont observés dans deux segments adjacents. On peut noter que, comme les ruptures candidates sont obtenues avec des règles locales appliquées au contenu de la mémoire cache, quand on génère des segments de textes qui ne contiennent pas un nombre suffisant de mots la rupture candidate est ignorée. Ce nombre a été fixé empiriquement en fonction de la méthode utilisée.

TAB. 4 – Méthode de repérage par le modèle cache et sélection par le modèle cache

θ	0,001	0,0014	0,0018	0,002	0,003
Rappel	0,3857	0,3797	0,3715	0,3618	0,2851
Précision	0,6492	0,6794	0,7394	0,7543	0,8141

TAB. 5 – Méthode de repérage systématique avec sélection par le modèle cache

N	1	5	10
Rappel	0,3958	0,4706	0,3726
Précision	0,4235	0,5197	0,4911

Les résultats sont dans les tables 4 et 5 en fonction de la méthode de détection des candidats qui a été utilisée. On remarque l'importance de la méthode de repérage lors de la phase de sélection, puisque les candidats choisis par la méthode à base de mémoire cache donnent de meilleurs résultats qu'avec la méthode systématique. On le voit notamment en comparant les cas où $\theta = 0,0018$ et $N = 10$. Pour la même valeur de rappel, la valeur de précision est nettement meilleure par le repérage à base de mémoire cache.

Distance entre le contexte gauche et le contexte droit

Dans cette méthode, nous recherchons une séquence optimale de candidats, selon un critère d'optimisation. L'avantage essentiel que propose cette méthode est que, contrairement à la précédente, elle ne nécessite pas de connaissances *a priori* des thèmes.

C'est une programmation dynamique dont l'automate est celui de la figure 2. On note r, le cas où la rupture candidate est effective, et c le cas où c'est une continuité.

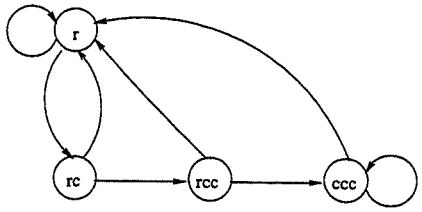


FIG. 2 – Automate de la méthode de sélection des candidats

Le treillis d'évaluation qui résulte de cet automate est présenté dans la figure 3. On note RP_i la rupture potentielle au i-ème segment. L'évaluation des

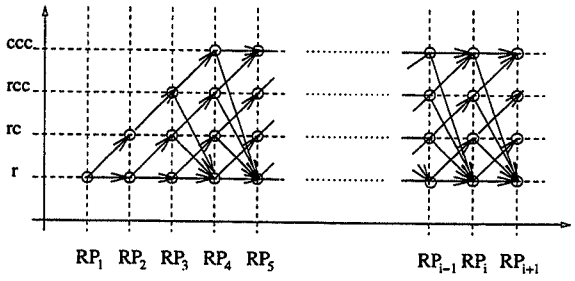


FIG. 3 – Treillis de la programmation dynamique selon l'automate de la figure 2

points de ce treillis s'effectue en 3 étapes. Dans un premier temps, on évalue la distance $d_i(G, D)$, distance de Kullback-Liebler entre le contexte gauche et le contexte droit, du i-ème candidat, où un nombre différent de segments peut être utilisé pour représenter le contexte gauche, en fonction de l'état pour lequel on calcule cette distance. Ceci est illustré par la figure 4. Ensuite, on analyse les variations de cette distance avec celles des états précédents potentiels :

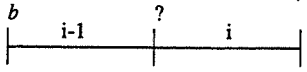
$$\Delta_i = d_i(G_i, D_i) - d_{i-1}(G_{i-1}, D_{i-1})$$

Enfin, on calcule les probabilités de continuité $P(c)$ et celles d'une rupture $P(r)$ telles que :

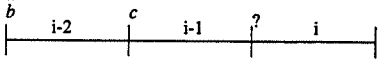
$$P(c) = \frac{\alpha}{1 + \exp^{-\Delta_i}}$$

$$P(r) = 1 - P(c)$$

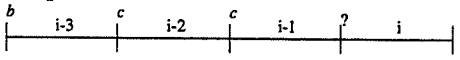
- état 1 : r. Le contexte gauche est représenté par un seul segment (i. e. la rupture candidate au $(i-1)$-ème segment est une rupture effective).



- état 2 : rc. Le contexte gauche contient les deux segments précédents.



- état 3 : rcc. Le contexte gauche contient les trois segments précédents.



- état 4 : ccc. Le contexte gauche contient les trois segments précédents.

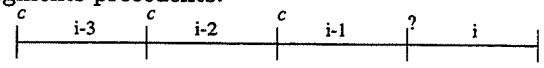


FIG. 4 – Différentes tailles de G, pour l'évaluation de la distance $d_i(G, D)$

L'état précédent que l'on choisit est celui dont $P(c)$ est la plus grande. On remarque que dans le cas où $\Delta_i = 0$ et $\alpha = 1$, on aura $P(c) = P(r) = 0,5$. Les résultats de cette méthode de sélection sont en tables 6 et 7.

TAB. 6 – Méthode de repérage par le modèle cache ($\theta = 0,002$), et sélection par la méthode à historique variable

α	3
Rappel	0,5477
Précision	0,5502

TAB. 7 – Méthode de repérage systématique et sélection par la méthode à historique variable

N	5	5	5	10
α	1	2	3	1
Rappel	0,6012	0,5705	0,5447	0,4284
Précision	0,1587	0,2607	0,3539	0,2473

2.3. *Synthèse des résultats*

La figure 5 donne la courbe de rappel et précision que l'on obtient avec l'ensemble des différentes méthodes utilisées. Ces résultats montrent que différentes stratégies doivent être utilisées selon les valeurs de rappel ou de précision que l'on cherche à obtenir.

Perspectives

Dans cet article, on a montré que l'on peut déterminer rapidement le thème d'un paragraphe. On a vu aussi que les modèles de langage résultant de la combinaison linéaire de modèles thématiques et d'un modèle général peuvent apporter des gains substantiels de perplexité. Ce résultat pourra par la suite être validé en dictée réelle. On a ainsi tous les éléments pour réaliser une adaptation en fonction des thèmes identifiés dynamiquement lors de la dictée. Le problème à résoudre est de disposer d'un corpus suffisant pour apprendre les modèles de langage thématiques. C'est pourquoi, nous avons développé des méthodes de segmentation en thèmes. Les meilleures méthodes proposées utilisent le modèle cache qui nécessite des connaissances préalables sur les thèmes. Il serait préférable que la segmentation thématique puisse être obtenue sans nécessiter ces connaissances *a priori*. Une partie de la solution est proposée par la méthode de sélection à historique variable. La suite de ce travail est de trouver des méthodes efficaces pour repérer les candidats qui, elles-aussi, ne nécessiteront pas de connaissances préalables sur les thèmes.

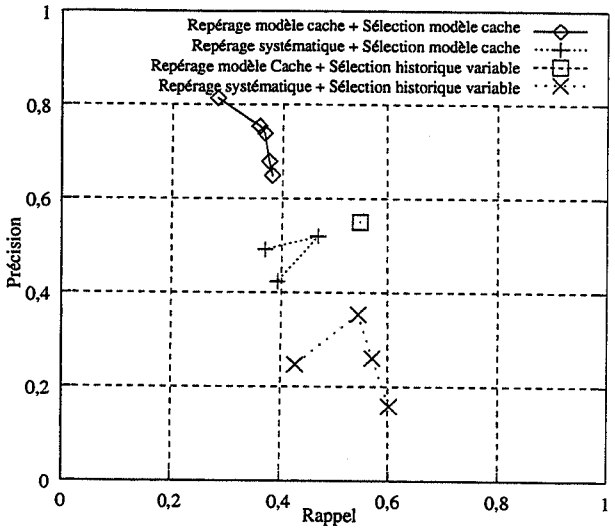


FIG. 5 – Résultats de la segmentation thématique (tables 4,5,6 et 7)

Références

[1] B. Bigi, R. De Mori, M. El-Beze, T. Spriet A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models *Special Issue on Fuzzy Logic in Signal Processing*, Signal Processing Journal, volume 80, numero 6, 2000.

[2] B. Bigi, R. De Mori, M. El-Beze, T. Spriet Detecting topic shifts using a cache memory *5th International Conference on Spoken Language Processing*, ICSLP-98, Sydney, Australia

[3] R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. In *IEEE Trans. Pattern anal. Machine Intell*, PAMI-12(6), pp 570–582, 1990.

[4] H. Li and K. Yamamishi. Document classification using a finite mixture model. Proc. of the *Conference of the Association for Computational Linguistics*, pp 39–47, Madrid, Spain, 1997.

[5] Peskin, S. Conolly, L. Gillick, S. Lowe, D. McAllaster, V. van Mulbregt, and S. Wegmann. Improvements in switchboard recognition and topic identification. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 303–306, Atlanta GA, 1996.

[6] R. Rosenfeld. The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation" Proc. of the *ARPA Spoken Language Technology Workshop*, pp 47–50, Austin, Texas, 1995.

Modélisation multi-bandes de la parole par champ de Markov

Guillaume Gravier, Marc Sigelle et Gérard Chollet

ENST-TSI & CNRS-URA 820
46, rue Barrault 75634 Paris Cedex 13
ggravier@infres.enst.fr, sigelle@tsi.enst.fr, chollet@tsi.enst.fr

Abstract

In this paper, an extension of the multi-band model that includes inter-band control of time asynchrony is described. The proposed model is based on the framework of Markov random fields. The law of the speech process is given by a parametric Gibbs distribution and a maximum likelihood parameter estimation algorithm is developed. This random field model is applied to isolated word recognition. It is shown that similar performances are obtained with the proposed model and with standard HMM techniques in the mono-band case. In a multi-band approach, results show that inter-band synchrony is an important parameter to take into account, in particular when dealing with noisy test signals.

1. Introduction

Les modèles de Markov cachés (MMC), utilisés dans la plupart des systèmes existants de reconnaissance de la parole (cf. par exemple [Jel98], chap. 2), donnent de bonnes performances. Cependant, cette modélisation souffre d'un certain nombre de limitations. En particulier, les performances des MMC se dégradent en présence de bruits additifs ou convolutifs et de distorsions, comme la réverbération. Pour éliminer les bruits convolutifs à variations lentes, comme les distorsions liées au canal téléphonique, on utilise en général la soustraction cepstrale ou le filtrage RASTA. Pour le traitement des bruits additifs, deux familles de solutions sont envisageables. Il est possible de rechercher une représentation du signal moins sensible au bruit que le représentation cepstrale ou, d'autre part, d'avoir un modèle statistique qui permette de traiter ce problème de manière efficace. Dans le premier cas, plusieurs représentations ont été proposées ces dernières années : spectre de modulation, TRAPS, analyse LDA, soustraction spectrale, etc. Dans le cadre de la modélisation statistique, des approches par MMC multi-bandes [HPT96] ont été proposées pour traiter le problème du bruit additif.

Dans l'approche multi-bandes, le signal est divisé en sous-bandes, chaque bande étant modélisée de manière indépendante à l'aide d'un MMC. Les scores partiels obtenus dans chaque bande sont ensuite recombinés. Entre deux points de recombinaison des scores, ce modèle repose sur une modélisation asynchrones des bandes. Par ailleurs, Tomlinson *et al.* ont montré l'intérêt de l'asynchronie entre les sous-bandes dans [TRM+97]. Cependant, l'hypothèse d'indépendance entre les sous-bandes dans l'approche multi-bandes ne parait pas réaliste. De plus, si les sous-bandes ne sont pas totalement synchrones, il semble cependant peu probable qu'elles soient totalement asynchrones. Notons également que l'asynchronie entre les bandes peut en partie être due au canal de transmission et ne présenter aucun intérêt pour la reconnaissance de la parole.

Nous proposons donc d'introduire des interactions entre les sous-bandes pour modéliser la synchronie spectrale en étudiant une modélisation de la parole basée sur les champs de Markov. Un tel modèle a précédemment été utilisé dans [GSC98b] pour modéliser la sortie d'un banc de filtre, les résultats obtenus étant décevants à cause de la trop grande variabilité de la représentation, et aussi en raison de l'absence d'algorithme d'estimation des paramètres. Dans cet article, nous proposons d'appliquer cette modélisation à une approche multi-bandes avec une représentation cepstrale du signal dans les bandes. Après avoir introduit le formalisme des champs Markoviens, nous décrivons dans un premier temps le modèle proposé. Nous rappelons également, section 3, l'algorithme d'estimation des paramètres et les stratégies de décodage en reconnaissance de mots isolés précédemment introduits dans [GSC98a]. Nous présentons finalement des résultats pour différentes architectures en sous-bandes et étudions le modèle en présence de bruit additif, avant de conclure.

2. Modélisation par champ Markovien

2.1. *Champ de Markov et distribution de Gibbs*

Un champ aléatoire X, défini sur un ensemble de sites (ou treillis) S, est Markovien si la probabilité d'observer une valeur en un site s ne dépend que d'un nombre fini de sites voisins V_s. L'ensemble des voisins est donné par un système de voisinage noté V. De manière formelle, la propriété Markovienne est donnée par $P[x_s|x_r\ \forall r \neq s] = P[x_s|x_r\ \forall r \in V_s]$.A un système de voisinage donné correspond un ensemble de cliques, une clique étant un ensemble de points du treillis mutuellement voisins. Le théorème de Hammersley-Clifford [Bes74] permet d'établir une correspondance entre un champ de Markov et un champ de Gibbs lorsqu'aucune réalisation de X n'est de probabilité nulle. La loi du champ X est alors

donnée par la distribution de Gibbs

$$P[x] = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} U_c(x) \right) \quad , \qquad (1)$$

où $U_c(x)$ correspond à un potentiel associé à la clique c, $\mathcal{C}$ désignant l'ensemble des cliques sur S pour V. La constante Z, appelée fonction de partition, assure que l'équation (1) définit une mesure de probabilité. Elle est donnée par une somme sur toutes les configurations x possibles. Les définitions précédentes montrent que la probabilité d'une configuration dépend d'un ensemble d'interactions locales, *i.e.* au niveau des cliques. On note aussi que plus l'énergie totale $U(x) = \sum U_c(x)$ est grande, moins la configuration x est probable.

Nous pouvons donc définir un modèle basé sur les champs de Markov en déterminant un système de voisinage et en définissant les potentiels pour chaque clique associé au système de voisinage choisi.

2.2. *Définition des potentiels*

Dans le MMC multi-bandes, la loi du processus (ou champ) caché X est donnée par les différents MMC en parallèle, la variable aléatoire $X_{t,k}$ ne dépendant que de $X_{t-1,k}$. On montre que cette relation a une équivalence bilatérale [Geo88] dans laquelle le voisinage du point (t,k) est donné par $\{(t-1,k),(t+1,k)\}$. Pour introduire une interaction entre les chaînes de Markov correspondant à chacune des bandes, on considérera le voisinage donné par

$$V_{t,k} = \{(t-1,k),(t+1,k),(t,l) \quad \forall l \neq k\} \ .$$

Un tel système de voisinage met en jeu deux types de cliques, *horizontales* et *verticales*, pour lesquelles nous allons définir un potentiel.

A partir d'études montrant qu'une chaîne de Markov est une distribution de Gibbs particulière [ZAZ91], nous définissons le potentiel associé aux cliques du type $\{(t-1,k),(t,k)\}$ par

$$U_{t,k}^{(h)} = \sum_{i,j} a_{ij}^{(k)} \, \delta(x_{t-1,k} = i) \, \delta(x_{t,k} = j) \quad , \qquad (2)$$

$\delta(x_{t-1,k} = i)$ prenant la valeur 1 si l'égalité est vérifiée et 0 sinon. Le paramètre $a_{ij}^{(k)}$, appelé poids de transition, est en fait homogène à $-\ln P^k(i,j)$, P^k étant la matrice de transition de la chaîne de Markov associée à la bande k.

En numérotant de 1 à N les états dans chaque bande et si l'on considère que deux bandes sont synchrones lorsque les changements d'états sont observés aux mêmes instants dans les deux bandes, alors la synchronie entre les bandes k et l peut-être modélisée en associant aux cliques $\{(t,k),(t,l)\}$ le potentiel défini par

$$U_{k,l}^{(v)} = f_{kl} \, |x_{t,k} - x_{t,l}| \ . \qquad (3)$$

En effet, lorsque le poids de synchronisation f_{kl} est grand, on favorise un comportement synchrone des bandes puisqu'alors $|x_{t,k} - x_{t,l}|$ est petit pour les configurations vraisemblables.

2.3. *Lois a priori et a posteriori*

En considérant les potentiels définis par les équations (2) et (3) et dans le cas d'un modèle à K bandes avec N états par bandes, la loi *a priori* de X est une distribution de Gibbs d'énergie totale

$$U(x) = \sum_{k=1}^{K} \sum_{i,j=1}^{N} a_{ij}^{(k)} \, \varphi_{ij}^{(k)}(x) + \sum_{k,l>k}^{K} f_{kl} \, \psi_{kl}(x) \ . \quad (4)$$

La fonction $\varphi_{ij}^{(k)}(x)$ compte le nombre de transitions de l'état i vers l'état j dans la bande k tandis que la fonction $\psi_{kl}(x)$ est l'écart cumulé entre les bandes k et l donné par

$$\psi_{kl}(x) = \sum_{t} |x_{t,k} - x_{t,l}| \ .$$

En associant à chaque état i dans chaque bande k une densité gaussienne, notée $g(\,.\,;\mu_i^{(k)},\sigma_i^{(k)})$, et sous l'hypothèse classique d'indépendance conditionnelle des données, la loi d'une observation $Y = y$ conditionnellement à $X = x$ est une distribution de Gibbs d'énergie

$$U(y|x) = -\sum_{t,k} \sum_{i} \left(\ln(g(y_{t,k};\mu_i^{(k)},\sigma_i^{(k)})) \, \delta(x_{t,k} = i) \right) \ .$$

On montre également que la distribution *a posteriori* de X connaissant $Y = y$ est donnée par une distribution de Gibbs d'énergie $U(x|y) = U(x) + U(y|x)$. Il est donc possible d'appliquer sur X les algorithmes classiques des champs de Markov pour les lois *a priori* et *a posteriori*.

3. Estimation des paramètres et algorithmes de décodage

3.1. *Estimation des paramètres*

Dans le cas de K sous-bandes, le modèle de champ de Markov est défini par l'ensemble des K matrices $(N \times N)$ de poids de transitions, $A^{(k)}$, la matrice $(K \times K)$ des poids de synchronisation, F, et les moyennes $\mu_i^{(k)}$ et variances $\sigma_i^{(k)}$ des gaussiennes. Nous proposons d'estimer simultanément l'ensemble de ces paramètres, noté θ, selon un critère du maximum de vraisemblance, en utilisant une procédure EM couplée à un algorithme de gradient pour l'étape de maximisation [Lan93].

Dans le cas d'une unique observation d'apprentissage, la fonction auxiliaire de l'algorithme EM est donnée par

$$\begin{aligned} Q(\theta,\theta^{(n)}) \ = \ & -\sum_{k} \sum_{i,j} a_{ij}^{(k)} E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y] - \\ & \sum_{k,l>k} f_{kl} E_{\theta^{(n)}}[\psi_{kl}(x)|y] - \ln Z_\theta - \\ & \sum_{t,k} \gamma_{t,k}(i) \ \ln g(y_{t,k};\mu_i^{(k)},\sigma_i^{(k)}) \end{aligned} \quad (5)$$

où Z_θ est la fonction de partition associée à (4), $\theta^{(n)}$ l'estimation courante des paramètres et $\gamma_{t,k}(i) = P_{\theta^{(n)}}[X_{t,k} = i | Y = y]$. En dérivant (5) par rapport à $a_{ij}^{(k)}$, on obtient l'équation de maximisation suivante

$$E_{\theta^{(n+1)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y] = 0 \qquad (6)$$

pour le poids de transition $a_{ij}^{(k)}$. L'équation (6) n'admettant pas de solution analytique, nous proposons de maximiser $Q(\theta, \theta^{(n)})$ en appliquant un pas de gradient pour obtenir une nouvelle estimation $\theta^{(n+1)}$ du paramètre, ce qui donne alors

$$a_{ij}^{(k)} \leftarrow a_{ij}^{(k)} + \frac{E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y]}{V_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)]} \quad . \quad (7)$$

Les espérances mises en jeu dans cette équation n'étant pas explicitement calculable, on les approxime à partir d'échantillons du champ X tirés suivants les lois *a priori* et *a posteriori*. Enfin, les paramètres $a_{ij}^{(k)}$ n'étant pas indépendants, cette méthode est appliquée en pratique à un vecteur regroupant les paramètres dépendants [GSC98a], la dénominateur dans (7) étant alors une matrice de covariance.

La réestimation des poids de synchronisation suit le même schéma que précédemment, ces paramètres étant supposés indépendants. Les formules de réestimation des paramètres des gaussiennes sont les mêmes que dans l'algorithme de Baum-Welch, les probabilités d'occupation des états, $\gamma_{t,k}(i)$, étant estimées à partir d'échantillons de X selon la loi *a posteriori*.

Les paramètres initiaux $\theta^{(0)}$, à l'exception des poids de synchronisation qui ne sont pas initialisés, sont obtenus à l'aide d'une stratégie itérative. Cette stratégie se base sur des estimateurs empiriques appliqués aux données complètes (x^*, y), x^* étant la configuration la plus vraisemblable *a posteriori*. Nous avons pu montré par des simulations que la procédure d'estimation des paramètres proposée donne de bons estimateurs.

3.2. Stratégie de décodage

La reconnaissance de mots isolés se base sur le calcul du score $p_w(y)$ pour chaque mot w du vocabulaire. Comme dans le cas des MMC, il est nécessaire de recourir à des approximations dans le calcul de ce score. Nous approximons $p_w(y)$ par le score sur les données complètes donné par

$$p_w(y) = pl_w(x^*) p_w(y|x^*) \quad . \quad (8)$$

Dans cette équation, la fonction $pl_w(x^*)$ correspond à la pseudo-vraisemblance de la configuration x^* et remplace la vraisemblance qui n'est pas calculable [Cha89]. La configuration x^* correspond à une estimation au sens du maximum *a posteriori* de X et peut-être déterminée à l'aide de l'algorithme ICM [Cha89] ou par recuit simulé. La différence entre ces deux algorithmes réside dans le fait que l'algorithme ICM converge rapidement mais n'est pas optimal, tandis que le recuit simulé converge vers un optimum global mais de manière plus lente.

4. Reconnaissance de mots isolés

4.1. Protocole expérimental

Le modèle par champ Markovien proposé est étudié en reconnaissance mono-locuteur de mots isolés sur de la parole téléphonique. Le vocabulaire est constitué de 10 mots courants. Un corpus contenant 100 occurrences de chaque mot du vocabulaire, prononcées par un même locuteur, a été extrait de la base de données PolyVar pour réaliser les expériences. Les cinquante premières occurrences de chaque mot sont utilisées pour l'estimation des paramètres des modèles, les 50 dernières étant réservées pour les tests.

Le signal de parole est divisé en K sous-bandes régulièrement réparties sur une échelle MEL, une représentation cepstrale du signal étant adoptée dans chaque bande. Le nombre N d'états par bande dépend du nombre de phonèmes composant le mot.

4.2. Résultats

Nous présentons dans la table 1 les résultats obtenus pour différents algorithmes d'estimation des paramètres et de décodages. Nous considérons une décomposition de la bande passante en 1, 3, 5 et 7 bandes représentées par, respectivement, 12, 5, 3 et 2 coefficients cepstraux. L'estimation ICM correspond à une simple initialisation des paramètres, x^* étant déterminé par l'algorithme ICM. Dans l'estimation ICM-EM, l'initialisation est suivie de 10 itérations de l'algorithme EM proposé. Dans les deux cas, le décodage peut-être basé sur l'algorithme ICM ou bien sur le recuit simulé (RS) pour la calcul de x^* dans (8). Enfin, pour comparaison, la première ligne du tableau correspond à une estimation par l'algorithme de Baum-Welch des paramètres des MMC de manière indépendante dans chaque bande. Dans ce dernier cas, pour le décodage V-ICM, la meilleure configuration x^* est déterminée par l'algorithme ICM initialisé par une segmentation obtenue en appliquant Viterbi dans chaque bande [GSC98b]. Le score est ensuite calculé comme précédemment à l'aide de l'équation (8).

Table 1: Taux de reconnaissance (en %) en fonction du nombre de bandes pour différentes techniques d'estimation des paramètres et de décodage.

estimation	scoring	b1c12	b3c5	b5c3	b7c2
BW	V-ICM	99.8	99.4	97.6	95.0
ICM	ICM	87.2	84.6	78.8	78.2
ICM-EM	ICM	88.6	80.6	75.4	76.0
ICM	RS	99.6	97.8	92.6	88.2
ICM-EM	RS	99.0	97.8	95.0	94.2

Les résultats mettent en évidence le problème du décodage basé sur l'algorithme ICM, ce dernier étant trop sensible aux conditions initiales. En dépit des différences dans les procédures d'estimation des paramètres entre les trois premières lignes du tableau, l'écart de performance entre le décodage V-ICM et ICM montre clairement les défauts de l'algorithme ICM. Dans le cas mono-bande (b1c12), le modèle de champ de Markov proposé donne des résultats similaires à ceux obtenus avec les MMC, pour un décodage à base de recuit simulé. La comparaison des différentes divisions en sous-bandes du signal montre que dans tous les cas, le taux de reconnaissance est inversement proportionnel au nombre de bandes. Ce résultat pourrait être expliqué par le fait que l'on

a une moins bonne représentation de la parole, le nombre de coefficients cepstraux utilisés dans chaque bande devenant de plus en plus faible. Cependant, lorsque l'on prend 5 coefficients cepstraux par bandes dans un modèle à 7 bandes, le taux de reconnaissance n'augmente que de manière marginal, passant de 95% à 96.4% dans l'approche par MMC indépendants. Dans le décodage par recuit simulé, on peut voir que l'estimation des poids de synchronisation par l'algorithme EM permet de limiter la baisse de performance. La modélisation de la synchronisation inter-bandes a donc une influence importante sur les performances du modèle. Ce résultat souligne la nécessité de disposer d'un bon modèle *a priori* du processus caché X lorsque l'observation devient plus variable. En effet, la variabilité de l'observation dans une bande croît avec le nombre de bandes. Dans ce cas, une meilleure modélisation du processus X permet une meilleure régularisation de la segmentation, et donc un meilleur taux de reconnaissance.

Pour étudier le comportement de l'approche proposée en présence de bruit additif, nous avons artificiellement ajouté un bruit aux données de test. Le bruit ajouté est coloré par un filtre de réponse impulsionelle $H(z) = 1/1 - 0.9z^{-1}$, concentrant ainsi l'énergie du bruit dans les basses fréquences. Les trois premières lignes du tableau 2 montrent les résultats obtenus pour 3 bandes, en appliquant une reconnaissance par MMC dans chacune des bandes pour différents rapports S/B. La quatrième ligne donne le taux de reconnaissance obtenu si l'on fusionne les scores obtenus avec les MMC dans chaque bande en faisant une moyenne. Enfin, la dernière ligne montre les résultats obtenus avec une approche par champs de Markov. Ces résultats montrent que dans ce cas, la fusion des scores par moyenne donne de mauvais résultats lorsque le rapport S/B augmente, les deux premières bandes étant fortement dégradées comme le montrent les taux de reconnaissance par bande. Le modèle de champ de Markov proposé donne de meilleurs résultats que la fusion par moyenne mais les performances obtenues sur la seule bande peu bruitée (bande #3) restent meilleures. Enfin, lorsque la synchronie entre les bandes n'est pas modélisée (*i.e.* $f_{kl} = 0$), le taux de reconnaissance à 30 dB n'est plus que de 44.8 % au lieu de 49.4 %, ce qui montre l'intérêt de la synchronie en présence de bruit.

Table 2: Taux de reconnaissance (en %) en fonction du rapport S/B dans une approche 3 bandes.

S/B (dB)	∞	30	20	10
bande #1	96.4	9.0	9.8	10.0
bande #2	94.4	17.4	17.4	14.2
bande #3	92.2	80.4	59.8	28.4
moyenne	99.6	28.0	13.2	10.4
ICM-GEM/RS	93.6	49.4	37.8	24.8

5. Conclusion

Dans cet article, nous avons proposé un modèle multi-bande dans lequel la synchronie entre les bandes est modélisée. Ce modèle, qui repose sur la théorie des champs de Markov, est étudié pour la reconnaissance de mots isolés. La méthode proposé donne des performances comparables aux MMC dans le cas mono-bande. Quelque soit l'approche choisie, les performances des systèmes baissent dans le cas de la parole non bruitée lorsque le nombre de bandes augmentent. Cependant, l'introduction d'un terme de synchronisation entre les bandes permet de limiter la baisse des performances. Enfin, en présence de bruit additif, le modèle proposé permet d'améliorer les performances par rapport à une approche par MMC avec une fusion par moyenne des scores partiels de chaque sous-bande.

Cependant, le modèle de synchronisation introduit dans notre approche est stationnaire dans le sens où les poids de synchronisation inter-bandes sont constants pour un segment donné. Bien que pratique, cette hypothèse de stationnarité de la synchronisation est fausse et un modèle plus complet devrait être étudié. Soulignons pour conclure que l'intérêt d'une modélisation par champs de Markov réside dans la souplesse de cette approche, de nombreux types d'interactions et de nombreuses familles de potentiels pouvant être envisagés.

Bibliographie

[Bes74] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statistical Soc.*, B-48:192–236, 1974.

[Cha89] Bernard Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747–761, 1989.

[Geo88] Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9 of *Studies in Mathematics*. de Gruyter, 1988.

[GSC98a] G. Gravier, M. Sigelle, and G. Chollet. Markov random field modeling for speech recognition. *Australian J. for Intelligent Info. Proc. Systems*, 5(4), 1998.

[GSC98b] G. Gravier, M. Sigelle, and G. Chollet. Toward Markov random field modeling of speech. In *ICSLP*, December 1998.

[HPT96] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *ICSLP*, Oct. 1996.

[Jel98] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[Lan93] Kenneth Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statistical Soc.*, 57(2):425–437, 1993.

[TRM+97] M. J. Tomlinson, M. J. Russel, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *ICASSP*, volume 2, pages 1247–1250, 1997.

[ZAZ91] Yunxin Zhao, Lee A. Atlas, and Xinhua Zhuang. Application of the Gibbs ditribution to hidden Markov modeling in speaker independent isolated word recognition. *IEEE Trans. on Signal Processing*, 39(6):1291–1298, 1991.

Traitement des mots hors-vocabulaire en compréhension de la parole

Caroline Bousquet-Vernhettes, Nadine Vigouroux et Guy Pérennou

IRIT / UMR CNRS 5505
118 route de Narbonne – Toulouse, France
Tél.: ++33 (0)5 61 55 66 11 poste 72 01 - Fax: ++33 (0)5 61 55 62 58
Mél : {bousquet, vigourou, perennou}@irit.fr - http://www.irit.fr/SSI/ACTIVITES/EQ_IHMPT/home.html

ABSTRACT

Conversational automatic inquiry systems must allow a large public to express themselves in a spontaneous way. From this point of view, robust spoken language understanding plays a crucial role for taking into account variability, various spoken and discourse phenomena and natural language ambiguities. In this paper we present the CACAO (Compréhension Automatique par segments Conceptuels Assistée par Ordinateur) environment based in a stochastic conceptual approach. In particular we will study how to take into account out-of-vocabulary words. Results obtained by our system will be presented.

1. INTRODUCTION

Comprendre les énoncés oraux est l'une des étapes importantes d'un serveur vocal interactif conversationnel en raison de leur caractère incertain et ambigu. Le processus de compréhension d'un dialogueur orienté par une tâche consiste à extraire le sens utile de l'énoncé par rapport à cette tâche. Ce sens inclut des informations référentielles et/ou des valeurs illocutoires [Pér96]. Confrontée à la variabilité de la parole —et spécialement lorsqu'il s'agit d'usagers qui s'expriment spontanément—, aux performances des systèmes de reconnaissance automatique de la parole (SRAP) utilisés en amont et aux ambiguïtés du langage naturel cette compréhension doit être **robuste**. En particulier il faut limiter les erreurs de compréhension dues à une mauvaise interprétation de mots dits hors-vocabulaire (HV), qu'il s'agisse de mots du vocabulaire mal reconnus par le SRAP ou de mots inconnus du modèle de langage (si ce dernier n'a pas une couverture linguistique suffisante).

Au niveau des SRAP, quelques laboratoires se sont préoccupés de ce type d'erreurs. Certains essaient de diminuer le taux des mots HV en augmentant tout simplement la taille du lexique [Add96] ou encore en tenant compte des formes fléchies ou des mots composés [Geu98]. Dans [You94] les auteurs essaient de détecter ces mots HV dans le but de déterminer leur sens et de les ajouter automatiquement dans le lexique afin de pouvoir ensuite les reconnaître et les comprendre. Dans [Gal96] les auteurs cherchent aussi à détecter les mots HV. Ceci est utilisé dans le système d'informations sur les horaires de train allemand, les modules de compréhension et de stratégies de dialogue ayant été prévus pour tenir compte des mots HV [Bor97].

En vue d'aborder de manière systématique la compréhension robuste de la parole au niveau sémantique et pragmatique, nous avons développé l'environnement CACAO (Compréhension Automatique par segments Conceptuels Assistée par Ordinateur) utilisant une modélisation stochastique de segments conceptuels.

Dans notre communication nous préciserons en premier lieu comment sont pris en compte les 'mots HV' dans notre modélisation conceptuelle ; nous présenterons ensuite l'environnement CACAO et en particulier nous expliquerons comment le paramétrer pour tenir compte des difficultés de compréhension dues à ces mots HV. Un scénario de deux tests (performances globales et tests sur les villes HV) et leurs résultats sont discutés.

2. MODÉLISATION CONCEPTUELLE ET MOTS HORS-VOCABULAIRE

Généralement les mots HV sont définis comme des mots qui sont inconnus de l'application parce qu'ils n'appartiennent pas au lexique de la compréhension. Cependant, la sortie du SRAP ne peut comporter que des mots connus par ce dernier (excepté dans certains cas où le SRAP indique simplement que le mot est inconnu [Gal96]). Par conséquent, le lexique de la compréhension étant en général le même que celui du SRAP, ces mots ne sont pas reconnus comme des mots HV mais confondus avec d'autres mots appartenant bien au lexique de la reconnaissance (et donc de la compréhension).

Ici, nous nous intéressons à un modèle de compréhension de la parole fondée sur la notion de segments conceptuels (SC). Elle a d'abord été introduite implicitement dans [Ord94] dans ce que les auteurs appellent graphe conceptuel. Cette notion a été explicitée et généralisée dans [Pér96]. Les SC sont définis comme des segments d'énoncés réalisant une combinaison d'actes référentiels et/ou illocutoires. Ces combinaisons doivent être attestées dans les pratiques langagières et les énoncés de la parole spontanée doivent, modulo des éléments assimilés à du bruit, être la concaténation de SC.

Par exemple, considérons les deux groupes de mots « aux environs de 16 heures » et « A 2 heures et demi du matin » : ils réalisent un (seul) acte référentiel concernant

le concept d'horaire et font partie de la même classe de SC 'horaire'.

Ainsi le langage de l'application est modélisable par un graphe conceptuel. Chaque classe de SC est représentée par une grammaire sous la forme d'un modèle de Markov caché dont les états émettent des mots. La classe de mots que peut émettre un état, présente en général, une parenté quant à leur rôle dans l'expression des actes référentiels et/ou illocutoires. Une telle classe peut ne pas avoir d'unité au plan syntaxique.

Tous les énoncés peuvent donc être décomposés en une suite de segments conceptuels comme dans l'exemple ci-dessous. Nous voyons donc qu'à chaque SC correspond un lexique qui est en fait une partie du lexique global du module de compréhension.

« Je voudrais aller à Paris demain vers 16 heures »

demande destination date horaire

Resituons les mots HV du point de vue de la compréhension : un mot sera considéré comme HV pour une classe de SC (respectivement pour une classe de mots) s'il n'appartient pas au lexique correspondant à cette classe de SC (respectivement à cette classe de mots).

Ainsi si nous considérons une application de demande d'informations sur les horaires de train, l'énoncé suivant « Je vais à Tours », peut être reconnu « Je vais à jour », en particulier si la ville de Tours, pour une raison quelconque, n'est pas connue par le système. Or le mot 'jour' sera considéré comme un mot HV pour le SC 'direction' ainsi que pour la classe de mot 'ville' de ce SC. Dans un tel cas, il serait intéressant de pouvoir identifier que l'usager parle d'une ville que le système de compréhension n'a pas su reconnaître, ce qui permettrait de mieux gérer la poursuite du dialogue.

3. PRÉSENTATION DE CACAO

CACAO est l'environnement de mise en œuvre de la compréhension basée sur les SC et le décodage stochastique. Dans cette approche, un même mot peut avoir des interprétations différentes selon le SC où il se trouve [Bou99b].

3.1 Fonctionnalités générales

CACAO se compose essentiellement de quatre modules correspondant aux quatre grandes étapes de notre processus de compréhension (figure 1) :

Le module de prétraitement permet de transformer la sortie de la reconnaissance de la parole sous la représentation adéquate pour le décodage conceptuel ; les mots considérés comme inutiles (hésitations, mots exprimant la politesse si on ne veut pas traiter cet aspect...) sont supprimés. Seuls les N meilleurs énoncés sont conservés.

Ensuite chaque énoncé est décomposé en une suite de segments conceptuels au moyen du décodage conceptuel. C'est à ce niveau qu'interviennent les connaissances linguistiques (lexique, graphes conceptuels) liées à l'application. Plusieurs algorithmes sont disponibles pour effectuer ce décodage et il est possible de tenir compte de l'historique du dialogue. Plusieurs types de paramètres permettent d'ajuster le modèle dans une tâche de compréhension. Deux d'entre eux permettent de prendre en compte les mots HV (ces paramètres seront détaillés dans la section suivante), les autres servent pour optimiser l'algorithme de décodage ou pour éventuellement traiter le contexte, mais nous n'en parlerons pas ici.

Le module de décision identifie la meilleure solution en tenant compte à la fois du score de la décomposition conceptuelle, de celui de la reconnaissance de la parole et éventuellement de l'historique du dialogue.

Cette solution est alors interprétée afin de fournir le sens de l'énoncé sous la forme d'un formulaire ou bien d'une structure de traits.

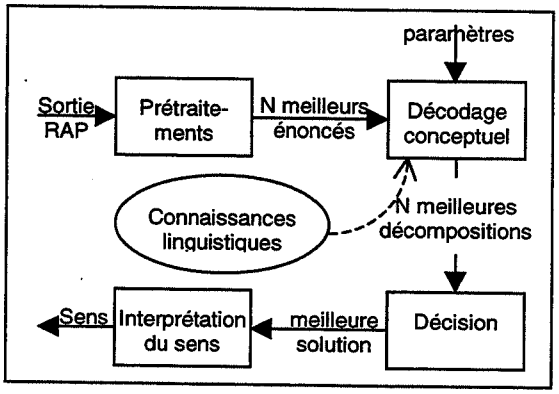


Figure 1 : Architecture de CACAO

3.2 Traitement des mots hors-vocabulaire

Comme nous l'avons indiqué plus haut, le module de décodage conceptuel peut être paramétré pour tenir compte des difficultés de compréhension liées aux mots HV.

Deux paramètres permettent d'ajuster les probabilités relatives au SC 'poubelle'[1]. Le premier permet de modifier le coût pour entrer dans le SC 'poubelle' et le second, le coût pour rester à l'intérieur de ce segment. Ceci permet de modifier facilement la probabilité d'interpréter un groupe de mots comme inutiles à la compréhension.

Deux autres paramètres permettent de modifier la probabilité de trouver un mot HV dans une classe de mots d'un SC. Le premier s'applique à toutes les classes de mots et permet de ne pas interdire totalement la présence d'un mot HV dans une classe (bien que cela soit peu probable). Le second paramètre ne s'applique qu'à

[1] C'est le segment qui récupère les mots inutiles pour la compréhension.

certaines classes définies à l'avance dans le modèle de langage. En effet nous considérons que certaines classes sont plus susceptibles d'accueillir des mots HV que d'autres. Considérons, par exemple le cas de la classe de mots 'ville' (regroupant des noms de villes) dans une application de demande d'information sur les horaires de train. Les villes connues par le système seront celles où il y a une gare mais l'usager peut demander d'aller dans une ville non desservie par le train et donc inconnue du lexique de l'application. Il est par conséquent plus probable qu'un mot HV appartienne à cette classe plutôt qu'à une autre. De plus cette classe émet des mots indispensables à la compréhension de l'énoncé.

4. PREMIERS RÉSULTATS

4.1 Présentation des tests

Nous avons choisi une application de demande d'informations sur les horaires de train de la SNCF. Notre laboratoire ayant participé au projet européen ARISE [Bag99], nous réutilisons les corpus obtenus pour tester notre modèle. Les tests ont tous été réalisés à partir des transcriptions orthographiques d'énoncés réels provenant de la plate forme DEMON [Pér98] développée à l'IRIT.

Dans le but d'étudier l'aptitude de notre module de compréhension à traiter des mots HV, nous avons réalisé deux types de tests (les valeurs des paramètres sont identiques pour les deux tests).

Test 1 : c'est en fait un test général puisqu'il contient n'importe quel type de mots HV, que ces derniers soient indispensables pour extraire le sens utile de l'énoncé ou non. Ils peuvent donc intervenir dans n'importe quelle classe de mots de n'importe quel SC. La majorité de ces mots ne font pas partie du lexique du processus de compréhension car l'apprentissage du modèle de langage n'est pas complet.

Test 2 : pour ce test nous avons choisi d'étudier le cas des villes inconnues, la probabilité de trouver un mot HV dans la classe de mot regroupant les noms des villes étant assez élevée.

4.2 Test général

Ce test a été réalisé sur un corpus total (noté CT) de 626 énoncés (n'appartenant pas à l'apprentissage) dont 80 contiennent un ou plusieurs mots HV (corpus CHV). Le modèle de langage de la compréhension comporte 30 classes de SC et a été appris sur 1873 énoncés, ce qui est peu mais suffisant pour voir l'influence des mots HV sur le processus de compréhension.

Nous remarquons que la grande majorité des énoncés mal découpés en SC contiennent un ou plusieurs mots HV. Cependant plus de 60% de ces énoncés sont tout de même bien étiquetés en SC (figure 2). Le taux de SC corrects (93.5 % pour CT) est similaire au taux d'énoncés corrects (93.6 %).

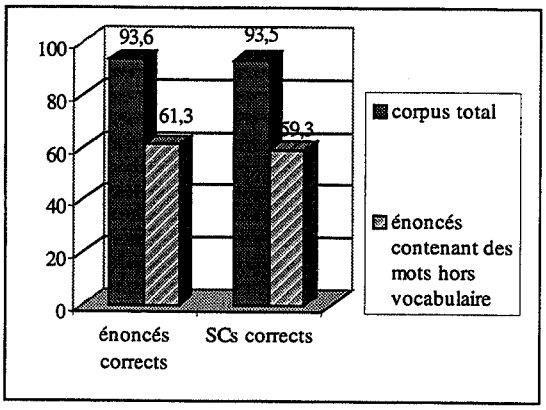


Figure 2 : Taux d'énoncés et de SC corrects – test 1

4.3 Tests sur les villes hors-vocabulaire

Les 1030 énoncés de ce corpus de test contiennent tous au moins une ville (767 énoncés contiennent une ville et 263 en contiennent deux voire trois ou quatre). Ce corpus comporte au total 4341 segments conceptuels, soit en moyenne 4.2 segments par énoncé. Ces villes ont toutes été remplacées par un mot HV. Toutefois, ce corpus ne contient pas d'énoncés très ambigus : à savoir, nous ne considérons pas les énoncés contenant une ville seule car une fois remplacée par un mot HV, on ne peut plus comprendre que l'on parlait d'une ville, à moins de tenir compte du contexte. Nous avons choisi de faire ce test sur des énoncés appartenant au corpus d'apprentissage ou bien à des énoncés qui sont bien compris, afin que les erreurs résultent uniquement de la présence du (ou des) mot(s) HV introduit(s) dans ces énoncés. Les résultats obtenus dépendent du choix de ce mot HV. Pour étudier cette influence nous avons réalisé cinq fois le même test avec à chaque fois un mot différent :

- Mot 1 ('truc') : Nous avons choisi un mot qui n'appartient pas au modèle de langage de la compréhension. Ce mot pourrait être l'étiquette OOV (Out Of Vocabulary) de la RAP décrite dans [Gal96].
- Mot 2 ('alors') : Nous avons choisi un mot appartenant au modèle de langage mais uniquement dans la classe du SC 'poubelle'.
- Mot 3 ('bon') : Ce mot appartient à la fois au SC 'poubelle' et aux SC de réponse (par exemple : « non, c'est bon »)
- Mot 4 ('sept') : Ce mot, contrairement aux deux précédents, ne se retrouve pas dans le SC de poubelle mais on le retrouve dans les segments de date et d'horaire (cas d'ambiguïté).
- Mot 5 ('jour') : Comme le mot 4, ce mot se trouve dans le SC de date mais il est bien moins fréquent.

Nous voyons ici (figure 3) que le taux d'énoncés corrects varie entre 73.5 % et 98.5 % selon le type du mot HV choisi. Le meilleur résultat est pour le mot 1 car ce mot n'appartenant pas au modèle de langage de la compréhension, le découpage en SC est moins ambigu

que dans les autres cas. En effet, toutes les erreurs de ce

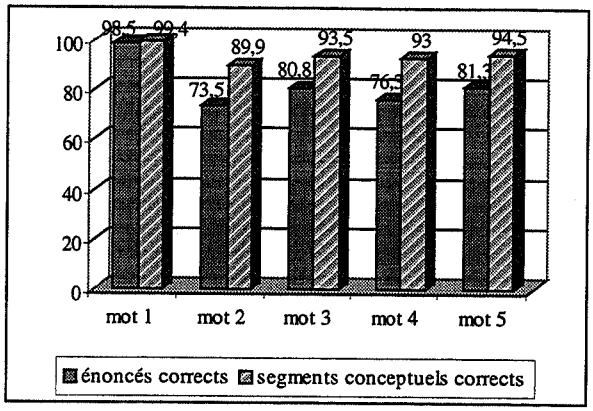


Figure 3 : Taux d'énoncés et de SC corrects – test 2

test sont dues à une confusion du SC contenant le mot HV, i.e. la ville, (segment de 'départ' ou de 'destination') avec le segment 'poubelle'. Nous retrouvons le même type d'erreur avec le mot 2 mais bien plus souvent que dans le premier cas car la probabilité de trouver ce mot dans le segment 'poubelle' est très élevée. Par contre, pour les trois autres mots, nous remarquons aussi des erreurs dues à la confusion avec le (ou les) SC auxquels ces mots appartiennent. En particulier le mot 4 ('sept') est souvent interprété comme appartenant aux SC de date ou d'horaire.

Les taux de SC corrects sont plus élevés que ceux au niveau des énoncés car la présence d'un mot HV influe peu sur les autres SC. Par exemple considérons la phrase « Je voudrais partir vers Paris à partir de **sept** demain dans la matinée » ; le mot 'sept' n'influe pas sur la compréhension de la ville de destination, de la date ou du moment du départ. Par contre, 'sept' risque d'être interprété comme un horaire et non comme une ville de départ inconnue.

5. Conclusion

Nous avons réalisé l'environnement CACAO permettant de mettre en œuvre la compréhension robuste de la parole spontanée dans le dialogue. La compréhension passe par une décomposition en segments conceptuels et utilise une modélisation markovienne multi-niveau. Les tests effectués montrent que malgré des données d'apprentissage réduites, notre modèle est relativement robuste aux difficultés soulevées par la présence de mots HV. Par la suite nous avons l'intention de tenir compte de l'historique du dialogue, ce qui aidera l'interprétation des mots HV. Ces tests sont encore trop limités car ils portent essentiellement sur les villes inconnues (bien que ce soit l'une des principales erreurs dans le type d'application que nous avons choisi) et il faudra les compléter dans le cadre d'une campagne d'évaluation des systèmes de dialogue.

Remerciements

Les auteurs remercient Martine de Calmès pour les transcriptions orthographiques des corpus.

Bibliographie

[Add96] Adda-Decker M., Adda G., Lamel L. et Gauvain J.L. (1996), « Developments in Large Vocabulary, Continuous Speech Recognition of German », ICASSP'96.

[Bag99] Baggia P., Kellner A., Pérennou G., Popovici C., Sturm J. et Wessel F. (1999), "Language Modelling and Spoken Dialogue Systems - the ARISE Experience", EUROSPEECH'99 Vol 4, pp. 1767-1770.

[Bou99a] Bousquet-Vernhettes C. (1999), « Compréhension Robuste de la Parole Spontanée par Segments Conceptuels », Atelier Thématique « Méthode Hybride TALN / TALP pour le Traitement Robuste de la Langue », TALN'99, pp. 51-60.

[Bou99b] Bousquet-Vernhettes C. (1999), « Stochastic Conceptual Model for Spoken Language Understanding », SPECOM'99, pp. 71-74.

[Bor97] Boros M., Aretoulaki M., Gallwitz F., Nöth E. et Niemann H (1997), « Semantic Processing of Out-of-vocabulary Words in a Spoken Dialogue System », EUROSPEECH'97, pp.1887-1890.

[Gal96] Gallwitz F., Nöth E. et Niemann H. (1996), « A Category Based Approach for Recognition of Out-of-Vocabulary Words », ICSLP'96, pp. 228-231.

[Geu98] Geutner P., Finke M. et Sheytt P. (1998), « Adaptative Vocabularies for Transcribing Multilingual Broadcast News », ICASSP'98.

[Oer94] Oerder M. et Aust H. (1994), « A realtime prototype of an autoamatic inquiry system », ICLSP'94, pp. 707-710.

[Pér96] Pérennou G. (1996), « Compréhension du Dialogue Oral – le Rôle du Lexique dans l'Approche par Segments Conceptuels », Lexique et Communication Parlée, GDR PRC, pp. 169-178.

[Pér98] Pérennou G., De Calmès M., Lavelle A. et Tronel R. (1998) "Un Système de Dialogue Oral Spontané pour l'Accès Téléphonique aux Informations d'Horaires de Train - Problème de Robustesse", Systèmes Complexes, Systèmes Intelligents et Interfaces, pp. 211-216.

[You94] Young S.R. (1994), « Recognition Confidance Measures : Detection of Misrecognitions and Out-of-vocabulary Words », ICASSP'94.

Équilibrage de charges dans un apprentissage parallèle pour la reconnaissance de la parole

E. M. Daoudi, A. Meziane, Y. O. Mohamed El Hadj

Université Mohammed I^{er}, Faculté des Sciences
Département de Mathématiques et d'Informatique
Laboratoire de Recherche en Informatique
60 000 Oujda, Maroc
E-mail:{mdaoudi, meziane, h.yahya}@sciences.univ-oujda.ac.ma

Abstract

In this paper, we propose a parallel technique of the training phase for automatic speech recognition using the centisecond Two Level Hidden Markov Model (TLHMM) which improves the load balancing in previous proposed parallel implementations. This technique is based on an efficient data distribution on processors taking into account not only the size of the vocabulary, but also the length of each sentence. In this manner, the idle time, induced by the load unbalancing between processors, will be reduced. The experimental results show that good performances can be obtained with this distribution.
Key words : automatic speech recognition, Markovian modeling, parallel processing, load balancing.

1. Introduction

A l'heure actuelle les systèmes classiques de la Reconnaissance Automatique de la Parole (RAP) les plus performants et les plus utilisés sont basés sur les modèles de Markov cachés (HMM) [4]. Mais ces modèles sont incapables de modéliser les phénomènes prosodiques de la parole, en particulier la durée des sons. Pour introduire ce paramètre, un nouveau modèle, Two Level Hidden Markov Model (TLHMM) [7], est développé. Cependant les algorithmes relatifs à ce modèle sont très coûteux en temps de calcul et en espace mémoire. Dans ce travail, nous proposons une implementation parallèle de la phase d'apprentissage d'un système de RAP utilisant la version centiseconde de ce modèle [4]. Dans cette étude, nous allons utiliser une technique de distribution de données permettant d'avoir un bon équilibrage de charges entre les processeurs.
A notre connaissance, peu de travaux relatifs à la parallélisation des systèmes de RAP ont été proposés dans la littérature [3, 5, 6].

2. Modèle séquentiel

2.1. *Structure du modèle*

Nous construisons de manière hiérarchique un réseau markovien de l'application en faisant appel à des connaissances linguistiques structurées à différents niveaux. Au niveau syntaxique, la phrase est vue comme concaténation de modèles de mots. Au niveau lexical, chaque mot est décrit par une séquence d'unités phonétiques et manipulé comme concaténation de modèles acoustiques. Au niveau acoustico-phonétique, un modèle de Markov caché est associé à chaque unité phonétique. En compilant toutes ces connaissances et en reliant les modèles des phrases par une entrée et une sortie communes, nous obtenons le modèle markovien global de l'application. L'entrée est un HMM représentant le silence du début des phrases, alors que la sortie correspond à un HMM modélisant le silence de fin des phrases.
Les paramètres des modèles markoviens sous-jacents sont appris sur un ensemble d'apprentissage formé par différentes prononciations de chaque phrase du vocabulaire. Ces paramètres sont :

- les probabilités des transitions entre les états, $(q_i)_{1\leq i\leq N}$, du modèle : $A = (a_{ij})_{1\leq i,j\leq N}$ où N est le nombre d'états de ce modèle.
- les distributions des probabilités qui régissent l'émission des observations acoustiques à partir des états : $B = \left(b_i(.)\right)_{1\leq i\leq N}$.
- les distributions des probabilités régissant les temps de séjour dans les unités phonétiques $\Delta = \left(\rho_k(.)\right)_{1\leq k\leq K}$ où K est le nombre d'unités phonétiques du vocabulaire modélisé.
- les probabilités initiales : $\Pi = (\pi_i)_{1\leq i\leq N}$.

Un TLHMM centiseconde sera noté $\lambda = (\Pi, A, B, \Delta)$. L'apprentissage de ces paramètres est effectué par une procédure de ré-estimation itérative qui consiste à déterminer, à partir d'un modèle initial λ_0 , un nouveau modèle λ_1 par rapport auquel la vraisemblance des observations conjointement au chemin optimal est maximale et à répéter la procédure jusqu'à satisfaction d'une condition d'arrêt.

2.2. *Algorithme de recherche du meilleur chemin*

Cet algorithme recherche le chemin optimal, au sens probabiliste, qui a vraisemblablement généré une suite d'observations $Y = y_1, ..., y_T$ dans un modèle λ tout en prenant en compte les durées de séjour dans les unités phonétiques alignées sur ce chemin. Il sauvegarde pour tous les états du modèle et pour toutes les observations acoustiques autant de chemins partiels que d'instants possibles d'entrée dans la dernière unité phonétique. Chacun de ces chemins est caractérisé par une grandeur d représentant le temps

passé dans l'unité phonétique courante. Ainsi, nous distinguons deux cas suivant la position dans l'unité phonétique courante :

- Pour $d \geq 2$, milieu de l'unité phonétique, nous avons la formule récurrente suivante :

  $$\delta_t(j,d) = \max_{i \in Pred_{\phi(j)}(j)} \Big(\delta_{t-1}(i,d-1) \times a_{ij}\Big) \times b_j(y_t)$$

  avec $Pred_{\phi(j)}(j)$ désigne l'ensemble des prédécesseurs de l'état q_j qui appartiennent à l'unité phonétique $\phi(j)$ dont provient l'état q_j. Pour mémoriser les états du chemin optimal, on utilise la variable $\psi_t(j,d)$ qui prend pour valeur, à chaque instant et pour tout $d \geq 2$, l'état qui a maximisé $\delta_t(j,d)$.

- Pour $d = 1$, début de l'unité phonétique, nous avons la formule récurrente suivante :

  $$\delta_t(j,1) = \max_{i \in \overline{Pred}_{\phi(j)}(j)} \left(\max_{1 \leq \overline{d} \leq d_{max}} \Big(\delta_{t-1}(i,\overline{d}) \times \rho_{\phi(i)}(\overline{d})\Big) \times a_{ij} \right) \times b_j(y_t)$$

  où d_{max} est la durée de séjour maximale dans une unité phonétique et $\overline{Pred}_{\phi(j)}(j)$ est l'ensemble des prédécesseurs de l'état q_j qui n'appartiennent pas à l'unité phonétique $\phi(j)$.
  Pour mémoriser les états du chemin optimal, dans ce cas, on considère la variable $\psi_t(j,1)$ qui a pour valeur, à chaque instant et pour $d = 1$, l'état qui a maximisé $\delta_t(j,1)$.

La valeur de δ_T, à l'instant T, permet de déterminer la probabilité d'émission le long du chemin optimal S^* :

$$Pr(Y,S^*) = \max_{i \in F} \left(\max_{1 \leq \overline{d} \leq d_{max}} \Big(\delta_T(i,\overline{d}) \times \rho_{\phi(i)}\Big) \right)$$

où F est l'ensemble des états finaux du modèle. Le couple (i_T^*, d^*) maximisant cette expression donne l'état final de ce chemin i_T^*. Les autres états sont obtenus par retour arrière à l'aide de la variable ψ_t de la façon suivante :
pour $t = T-1$ à 1 faire :
- Si $d^* = 1$ alors $(i_t^*, d^*) = \psi_{t+1}(i_{t+1}^*, 1)$
- Sinon $d^* = d^* - 1$ et $i_t^* = \psi_{t+1}(i_{t+1}^*, d^*)$

où d^* est initialisé à 1.

2.3. Complexité de l'apprentissage

Dans ce paragraphe, nous donnons une estimation du coût de l'apprentissage de chaque phrase du vocabulaire.
D'après la construction hiérarchique du réseau, chaque unité phonétique ϕ_τ est représentée par un modèle markovien de type gauche-droite. Le nombre d'état de ϕ_τ sera noté N^{ϕ_τ}.
Nous déterminons d'abord le nombre d'opérations flottantes (*flops*), qui sera noté $T_{cal}^{\phi_\tau}$, effectué dans une unité phonétique ϕ_τ relativement à l'algorithme de recherche du meilleur chemin.
• Pour d = 1, nous calculons la grandeur $\delta_t(j,1)$.
Le calcul de la loi acoustique $b_j(y_t)$ nécessite un nombre de *flops* fixe que l'on note C_1^{te}. De même, la loi phonétique $\rho_{\phi(i)}(\overline{d})$ nécessite un nombre de *flops* fixe que l'on note C_2^{te}. La quantité $\max_{1 \leq \overline{d} \leq d_{max}} \Big(\delta_{t-1}(i,\overline{d}) \times \rho_{\phi(i)}(\overline{d})\Big)$ demande $(1 + C_2^{te}) d_{max}$ *flops* et $d_{max} - 1$ comparaisons, au total $(2 + C_2^{te}) d_{max} - 1$ *flops* si nous supposons qu'une comparaison est équivalente à une opération flottante. Pour t et j fixent, nous obtenons

$$\Big(\big(2 + C_2^{te}\big) d_{max} + 1\Big) Card\Big(\overline{Pred}_{\phi(j)}(j)\Big) + C_1^{te} \quad flops$$

où $Card\Big(\overline{Pred}_{\phi(j)}(j)\Big) = np$, désigne le nombre moyen d'unités phonétiques directement attachées à l'unité phonétique courante. En faisant la somme sur t et j, nous obtenons

$$N^{\phi_\tau} \left(\Big(\big(2 + C_2^{te}\big) d_{max} + 1 \Big) np + C_1^{te} \right) T \quad flops$$

• Pour $d \geq 2$, nous calculons la grandeur $\delta_t(j,d)$. En utilisant le même raisonnement précédent, nous obtenons $2 \times Card\Big(Pred_{\phi(j)}(j)\Big) + C_1^{te}$ *flops* pour t, j et d fixent. Notons que pour $j = 2$ à N^{ϕ_τ}, $Card\Big(Pred_{\phi(j)}(j)\Big)$ est au plus égale à j dans le modèle de l'unité phonétique courante (modèle gauche-droite) et $Card\Big(Pred_{\phi(j)}(1)\Big) = 1 + np$. Donc nous avons :

$$\sum_{t=1}^{T} \sum_{d=2}^{d_{max}} \sum_{j=1}^{N^{\phi_\tau}} \Big(2 \times Card\Big(Pred_{\phi(j)}(j)\Big) + C_1^{te}\Big) \quad flops$$

Ce qui donne $(d_{max} - 1)\Big(2np + N^{\phi_\tau}(N^{\phi_\tau} + C_1^{te} + 1)\Big)T$.
Par conséquent :

$$T_{cal}^{\phi_\tau} \simeq \Big(2(d_{max} - 1)np + N^{\phi_\tau}\Big((d_{max} - 1)N^{\phi_\tau} + C^{te}\Big)\Big)T$$

où $C^{te} = \Big(C_1^{te} + np(2 + C_2^{te}) + 1\Big)d_{max} + np - 1$
Pour l'apprentissage d'une phrase, l'algorithme de recherche du meilleur chemin est utilisé dans le sous-réseau de cette phrase. Donc, si ϵ_i est le nombre d'unités phonétiques du modèle de la phrase numéro i, alors le nombre de *flops* nécessaire à l'apprentissage de la $j^{\text{ième}}$ prononciation de cette phrase est :

$$\sum_{i=1}^{\epsilon_i} \Big(2(d_{max} - 1)np + N^{\phi_\tau}\Big((d_{max} - 1)N^{\phi_\tau} + C^{te}\Big)\Big) T_i^j$$

$$\simeq \epsilon_i \Big(2(d_{max} - 1)np + \overline{N}_i^{\phi_\tau}\Big((d_{max} - 1)\overline{N}_i^{\phi_\tau} + C^{te}\Big)\Big) T_i^j$$

où $\overline{N}_i^{\phi_\tau} = \frac{\sum_{\tau=1}^{\epsilon_i} N^{\phi_\tau}}{\epsilon_i}$ est le nombre moyen d'états par unité phonétique relativement à la phrase numéro i et T_i^j le nombre des observations de cette prononciation.
En remarquant que $N_i = \epsilon_i \overline{N}_i^{\phi_\tau}$ représente le nombre d'états du sous-réseau de la phrase numéro i, nous obtenons une estimation du coût de l'apprentissage de n_i prononciation de cette phrase :

$$\sum_{j=1}^{n_i} \Big(2\epsilon_i(d_{max} - 1)np + N_i\Big((d_{max} - 1)\overline{N}_i^{\phi_\tau} + C^{te}\Big)\Big) T_i^j$$

$$= \Big(2\epsilon_i(d_{max} - 1)np + N_i\Big((d_{max} - 1)\overline{N}_i^{\phi_\tau} + C^{te}\Big)\Big) T_i$$

où $T_i = \sum_{j=1}^{n_i} T_i^j$ est le nombre total des observations de toutes les prononciations de la phrase numéro i.

3. Parallélisation

Nous considérons une architecture à mémoire distribuée composée de p processeurs, notés ($P_i)_{0 \leq i \leq p-1}$. Chaque processeur possède sa propre mémoire et communique avec les autres via un réseau d'interconnexion.
Dans [1] nous avons étudié une parallélisation du TLHMM centiseconde basé sur la duplication du réseau et dans [2], nous avons donné une autre stratégie de parallélisation, pour les HMMs, basant sur la distribution du réseau. A travers ces travaux nous avons constaté que la manière d'affecter les phrases à apprendre aux processeurs joue un rôle très important dans les performances des algorithmes proposés. Ceci est dû essentiellement à la différence de longueurs de ces phrases. Dans ce travail, nous proposons une amélioration du [1] par la mise en œuvre d'une technique de distribution de données qui tient compte non seulement de la taille du vocabulaire mais aussi du coût de l'apprentissage de chaque phrase.

Stratégie de parallélisation : le réseau global de l'application est affecté à chaque processeur, alors que le vocabulaire est uniformément distribué sur les processeurs. Cette distribution peut être faite de manière aléatoire en affectant à chaque processeur $\frac{m}{p}$ phrases, ou m est la taille du vocabulaire, indépendamment de leurs longueurs. Mais pour que le travail des processeurs soit équilibré, il faut tenir en compte, lors de la distribution, du coût de l'apprentissage de chaque phrase.
Selon l'étude de la complexité, le coût de l'apprentissage de la phrase numéro i est donné par :

$$\simeq \left(2\epsilon_i (d_{max} - 1)np + N_i \Big((d_{max} - 1)\overline{N}_i^{\phi_\tau} + C^{te} \Big) \right) T_i$$

Ce coût est calculé d'abord pour chaque phrase du vocabulaire puis, il est stocké par ordre décroissant dans un vecteur C. Les phrases sont ensuite affectées aux processeurs selon leurs coûts en utilisant une permutation circulaire sur les processeurs. Par exemple, la distribution de 6 phrases sur 3 processeurs est la suivante :

$C(0)$	$C(1)$	$C(2)$	$C(3)$	$C(4)$	$C(5)$
P_0	P_1	P_2	P_2	P_0	P_1

Une fois le vocabulaire est distribué, chaque processeur effectue l'apprentissage sur un corpus local de $\frac{m}{p}$ phrases dont chacune est prononcée n_i fois. Les informations résultantes de l'apprentissage de chaque phrase locale ne sont pas directement intégrées dans le réseau, mais plutôt combinées et stockées localement en vue de ré-estimer le réseau global ensuite. Un échange total entre tous les processeurs est ensuite effectué pour réactualiser le réseau global. Pendant cette phase de communication, qui est de type *all-to-all*, le processeur peut anticiper la ré-estimation des probabilités de transitions associées aux arcs formant les modèles de leur phrases locales. En effet, les modèles des phrases sont séparés entre eux et sont seulement reliées par les modèles de silences qui représentent l'entrée et la sortie communes du réseau global. Cette séparation entraîne l'indépendance des probabilités des arcs mais pas des lois acoustiques et phonétiques associées aux unités phonétiques, car ces dernières sont regroupées et modélisées par type d'unités phonétiques. Pour cette raison nous retardons la mise à jour des grandeurs caractérisant les lois jusqu'à rassemblement des données rendues par les différents processeurs. Aussi la ré-estimation des modèles de silence ne sera faite qu'après la terminaison de la communication.

Temps de calculs : si nous supposons que le coût de l'apprentissage ne dépend pas de la phrase à apprendre, alors le temps de l'apprentissage des phrases locales est égale au temps de l'apprentissage séquentiel divisé par p, grâce à la distribution du vocabulaire.

Temps de communications : pour les communications, le temps de la procédure *all-to-all* dépend de l'architecture d'interconnexion des processeurs. Sans perdre de généralité, nous supposons que les processeurs sont directement connectés et que chacun d'entre eux peut utiliser simultanément tous ses liens de communications. Dans ce cas le temps de communications, pour un modèle linéaire, de la procédure *all-to-all* est donné par $\beta + L\tau$ où β est le temps d'initialisation, τ est le temps de transfert d'une donnée et L est la taille du message à communiquer. La difficulté majeure qui se pose pour l'évaluation du temps de communication réside dans la détermination théorique de L. En effet, les messages sont formés par des informations concernant l'utilisation des transitions et des lois acoustiques relatives aux chemins optimaux associés aux phrases apprises. Or, ces informations sont difficiles à déterminer de manière exacte. De ce fait, nous allons prendre le nombre maximum de lois et de transitions possibles sur les chemins optimaux pour avoir une borne supérieure du temps de communications. Si la $j^{\text{ième}}$ prononciation de la phrase i à apprendre est composée de T_i^j observations alors le chemin optimal associé à cette phrase contient au plus $min(T_i^j, 2N_i - 1)$ transitions distinctes, où N_i est le nombre d'états du sous réseau de la phrase i, et ϵ_i^j unités phonétiques. Une transition est caractérisée, en plus de la loi acoustique qui lui est associée, par certains paramètres dont le nombre sera noté C_3^{te}. Les lois acoustiques que nous utilisons sont des multi-gaussiennes de vecteurs moyens et de matrices de covariances diagonales de taille μ tandis que les lois phonétiques sont de simples gaussiennes. Il s'ensuit que l'apprentissage de la $j^{\text{ième}}$ prononciation de la phrase i génère un message de taille $\simeq \Big(C_3^{te} + 2\mu\Big) min(T_i^j, 2N_i - 1) + 2\epsilon_i^j$. Donc le message obtenu par l'apprentissage d'un corpus local de $\frac{m}{p}$ phrases est :

$$\simeq \sum_{i=1}^{\frac{m}{p}} \sum_{j=1}^{n_i} \left(\Big(C_3^{te} + 2\mu\Big) min(T_i^j, 2N_i - 1) + 2\epsilon_i^j \right)$$

Par suite, le temps de communications pour une itération, est :

$$\simeq \beta + \left(\sum_{i=1}^{\frac{m}{p}} \sum_{j=1}^{n_i} \left(\Big(C_3^{te} + 2\mu\Big) min(T_i^j, 2N_i - 1) + 2\epsilon_i^j \right) \right) \tau$$

4. Expérimentations

Les évaluations sont faites sur un vocabulaire composé de 20 phrases dont chacune est prononcée une seule fois par chacun des 6 locuteurs pour former le corpus de l'apprentissage. Nous nous sommes limités à 20 phrases, puisque nous travaillons avec la technique de duplication du réseau qui limite la capacité de stockage sur chaque processeur (le même réseau est dupliqué sur tous les processeurs). Le volume de données à traiter peut être doublé en travaillant avec la technique de distribution du réseau [2] (le même réseau est dustribué sur tous les processeurs). Cette dernière technique est en cours de developpement pour le TLHMM centiseconde.
Ce vocabulaire est extrait de la base de données BDSONS (Base de Données des SONS français) qui est enregistrée au CNET Lannion avec une fréquence d'échantillonage de 16 khz. L'analyse acoustique que nous avons effectué sur ces données traite des blocs de signal de tailles fixes (32ms) avec recouvrement de moitié des blocs successifs. Chaque bloc est ensuite représenté par un vecteur spectral de 9 composantes, les 8 premiers coefficients cepstraux MFCC et l'énergie de ce bloc. Les programmes de cette partie sont rédigés en Matlab. Nous avons utilisé le pseudo-diphone comme unité de base pour construire le réseau markovien global de notre application. Chaque pseudo-diphone est modélisé par un HMM élémentaire avec des lois acoustiques multi-gaussiennes de vecteurs moyennes de taille 9 et de matrices de covariances diagonales. Les programmes parallèles développés sont implementés sous l'environnement de programmation PVM (Parallel Virtual Machine) sur la machine parallèle à mémoire distribuée TN310 composé de 32 nœuds de calculs dont chacun dispose d'un transputer T9000 pour le calcul et une mémoire de 8 Mo pour le stockage des variables locales.

Table 1: Temps de calculs moyen d'une itération de l'apprentissage pour la distribution aléatoire.

Processeurs	P_0	P_4	P_9
Temps	43.14	27.93	15.30

Table 2: Temps de calculs moyen d'une itération de l'apprentissage pour la distribution étudiée.

Processeurs	P_0	P_4	P_9
Temps	33.32	24.27	19.05

Table 3: Temps d'exécution moyen d'une itération de l'apprentissage.

Distrib.	Séq.	Parallèle			
nb. proc	1	2	4	5	10
aléatoire	189.97	161.40	98.84	83.85	43.57
étudiée	189.97	130.83	65.42	54.83	33.63

Dans la table 1 nous donnons les temps, en seconde, de calculs moyens d'une itération de l'apprentissage pour la distribution aléatoire obtenus par les processeurs, le plus rapide, l'intermédiaire et le plus lent. Ces données montrent un grand déséquilibre de charges entre les processeurs. Dans la table 2 nous donnons les temps, en seconde, de calculs moyens d'une itération de l'apprentissage pour les mêmes processeurs mais avec la distribution étudiée. Nous remarquons que le phénomène de dés-équilibrage de charges est réduit comparé à la première distribution. Dans la table 3, nous rapportons les temps d'exécution moyens, en seconde, d'une itération de l'apprentissage relativement aux deux distributions pour différents nombre de processeurs. Il montrent l'apport du parallélisme dans le domaine de la RAP. Ces résultats expérimentaux montrent que les meilleures performances sont obtenues pour la deuxième distribution ce qui est en accord avec l'analyse théorique.

5. Conclusion

Dans ce travail nous avons proposé une implementaion parallèle pour la phase d'apprentissage d'un système de RAP utilisant le modèle TLHMM centiseconde basé sur une duplication du réseau sur tous les processeurs. Dans cette implementation, deux stratégies de distributions de données ont été utilisées. La première stratégie (distribution aléatoire) consiste à affecter les phrases aux processeurs de manière aléatoire, tandis que la deuxième tient compte du coût de l'apprentissage de chaque phrase lors de la distribution. Les résultats expérimentaux montrent que les meilleures performances sont obtenues avec la seconde stratégie (distribution étudiée) et que le phénomène de dés-équilibrage de charges entre processeurs est largement réduit. Actuellement, nous sommes entrain d'étudier une nouvelle stratégie de distribution de données qui se base sur le coût global de l'apprentissage dans chaque processeur et non pas sur le nombre des phrases traitées par le processeur.

Bibliographie

[1] E.M. Daoudi, A. Meziane, Y.O. Mohamed El Hadj, "Parallel training for the automatic speech recognition using the centisecond TLHMM model", In Proceedings of ACIDCA'2000, pages , Tunisia 2000.

[2] E.M. Daoudi, A. Meziane, Y.O. Mohamed El Hadj, "Parallel HMM model for automatic speech recognition", Research Report, Faculty of Sciences of Oujda, LaRI, submitted for publication.

[3] M. Fleury, A. C. Downton, A. F. Clark, "Parallel Structure in an Integrated Speech-Recognition Network", In Proceedings of EUROPAR'99, pages 995-1004, 1999.

[4] A. Meziane, "Introduction de la durée des sons dans un modèle de Markov caché au niveau supra segmental", Thèse de doctorat d'état, Université Oujda, Avril 1997.

[5] H. Noda, M. N. Shirazi, "A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM", In Proceedings of ICASSP'94, pages I-597 - I-600, 1994.

[6] S. Phillips, A. Rogers, "Parallel Speech Recognition", In Proceedings of EUROSPEECH'97, 1997.

[7] N. Suaudeau, "Un modèle probabiliste pour intégrer la dimension temporelle dans un système de reconnaissance de parole". Thèse de doctorat de 3° cycle, Université de Renne I, Mars 1994.

Etudes comparatives des robustesses au bruit de l'approche 'Full Combination' et de son approximation

Astrid Hagen[§] *et Hervé Glotin*[§‡]

§ IDIAP, Martigny, Suisse
‡ ICP, Grenoble, France
hagen,glotin@idiap.ch - glotin@icp.fr

ABSTRACT

Sub-band based ASR aims to use reliable information from uncorrupted sub-bands only. For this one can select the best combination over all 2^d combinations of d sub-bands, without any need of independence assumption between sub-bands. We show that this approach can actually be set up by a combination function called the "Full Combination (FC)", constituting the fullband *posteriors* decomposed into a sum of weighted *posteriors* from 2^d experts. This approach is itself often not realizable as it encorporates training of too many experts. An approximation can estimate the *posteriors* for each combination based on the d sub-bands only, with a weak sub-band independence assumption. Tests under equal weights comparing FC and its approximation on band-limited noise show that for these noises the approximation performs as well as and sometimes better than FC, and both better than the baseline fullband system.

1. INTRODUCTION

La RAP (Reconnaissance Automatique de la Parole) multi-bandes exploite la redondance spectrale dans le but d'augmenter sa robustesse à l'inadéquation des données tout en faisant un minimum d'hypothèses sur le bruit interférant [Dup00, Cer99, Glo00]. Nous verrons que les experts des bandes non bruitées fournissent suffisamment d'information pour permettre un décodage robuste. Dans ce papier nous développons le modèle *Full Combination* (FC) qui s'inscrit dans le paradigme multi-bande de la RAP et dont les relations avec la perception humaine de la parole sont reprises dans [MHGB00]. Des expériences précédentes en RAP ont montré que le traitement indépendant des sous-bandes peut faire chuter les performances en parole claire. Une alternative consiste à travailler sur les estimations phonétiques des 2^d combinaisons (ou flux) des d sous-bandes, (y compris le flux vide correspondant aux probabilités *a priori*). Dans une première approche ces estimations ont été calculées pour chaque flux puis sélectionnées une à une [HTP96]. Dans notre approche FC, leurs estimations sont pondérées et sommées. Comme l'entraînement d'un expert par combinaison, soient 2^d experts pour d sous-bandes, est rapidement irréalisable, il est préférable de travailler avec des approximations de ces combinaisons. Nous montrons alors qu'il est possible d'obtenir des résultats similaires ou meilleurs au modèle FC avec son approximé «AFC», que nous comparerons sur des bruits de bande(s) idéaux, stationnaires centrés sur une des d sous-bandes, ou changeant de fréquence centrale toutes les 125ms. Après la description de la théorie et de l'implémentation des approches FC et AFC, les paramètres des systèmes hybrides HMM/ANN (ANN pour réseaux neuronaux ou Artifical Neural Network) ainsi que les données de test sont présentés. Puis suivent les expériences et leurs discussions.

2. LES MODÈLES FC ET AFC

2.1. L'approche «Full Combination»

Les systèmes multi-bandes pour la RAP décomposent le domaine spectral en plusieurs sous-bandes, qui sont traitées indépendemment, et dont les paramètres caractéristiques x sont passés aux reconnaisseurs correspondants. Les probabilités *a posteriori* $P(q_k|x)$ des sous-bandes sont recombinées dans le processus de reconnaissance. Dans notre approche les 2^d flux des combinaisons des d sous-bandes sont intégrés suivant ces évènements j_{propre} collectivement exhaustifs et mutuellement exclusifs : «la j^{ieme} combinaison de sous-bande est le flux qui produit la meilleure reconnaissance parmi tous les flux possibles». Considérant que les données bruitées hors du j^{ieme} flux propre sont négligeables dans l'estimation des probabilités *a posteriori* [LC97, MHGB00], nous posons $P(q_k|x, j_{propre}) \simeq P(q_k|x_j)$, où x_j est le vecteur acoustique du j^{ieme} flux. Nous avons alors:

$$P(q_k|x) \simeq \sum_{j=1}^{2^d} P(j_{propre}|x) P(q_k|x_j) \qquad (1)$$

Les probabilités dénotant les données claires $P(j_{propre}|x)$ dans (1) peuvent être estimées de différentes façons comme cela est démontré dans [BG99, HMB99, MHGB00, Glo00]. Pour les expériences présentées dans cet article, les $P(j_{propre}|x)$ sont équiprobables, ce qui est déjà fort intéressant en terme de robustesse comme nous allons le montrer.

Dans l'approche FC les termes $P(q_k|x_j)$ sont donnés en sortie du réseau de neuronne qui est entraîné et testé sur les paramètres acoustiques x_j. Dans l'approche AFC les termes $P(q_k|x_j)$ sont estimés à partir des sorties des ANN relatifs uniquement aux sous-bandes contenues dans la j^{ieme} combinaison.

2.2. *L'Approximation du FC : l'AFC*

Pour éviter l'entraînement de 2^d ANNs on peut estimer les probabilités $P(q_k|x_j)$ des combinaisons en utilisant uniquement les probabilités $P(q_k|x_i)$ issues des observations $x_{i,i\in\{1..d\}}$ des d sous-bandes qui composent cette combinaison (on notera J cet ensemble de sous-bandes, de cardinal $|J|$). Ce modèle ne requiert qu'une hypothèse d'indépendance des observations des sous-bandes conditionnellement à chaque classe phonétique, hypothèse plus faible que l'indépendance absolue [MHGB00]. Nous avons alors $P(x_j|q_k) \simeq \prod_{i\in J} P(x_i|q_k)$, donc

$$P(q_k|x_j)\frac{p(x_j)}{p(q_k)} \simeq \prod_{i\in J} P(q_k|x_i)\frac{p(x_i)}{p(q_k)} \qquad (2)$$

$$P(q_k|x_j) \simeq \frac{\prod_{i\in J} P(q_k|x_i)}{p^{|J|-1}(q_k)} \cdot \Theta \qquad (3)$$

avec $\Theta = \frac{\prod_{i\in J} p(x_i)}{p(x_j)}$, qui disparaît par normalisation sur toutes les classes phonétiques pour obtenir des estimations telles que : $\sum_k P(q_k|x_j) = 1$.

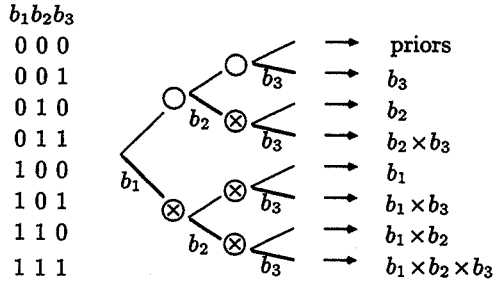


Figure 1: La valeur obtenue dans chaque feuille de l'arbre correspond à la multiplication des valeurs des branches en gras du chemin parcouru. On observe que le nombre de multiplications dans l'arbre (4) est inférieur à celui indiqué dans la colonne de droite (5), où le calcul est effectué indépendamment pour chaque combinaison. Cette différence augmente proportionnellement au nombre de bandes considérées.

Le calcul de l'approximation des probabilités $P(q_k|x_j)$ pour les 2^d combinaisons j peut se faire efficacement avec une procédure récursive qui réutilise les multiplications des composantes partagées par plusieurs combinaisons, au lieu d'effectuer un calcul indépendant pour chaque J. Ceci réduit considérablement le nombre de calcul lorsque le nombre de bandes considérées est grand, car par trame et par phonème 2^d–d–1 multiplications sont nécessaires au lieu de $d\times 2^{d-1}$–2^d+1, ce qui apporte une réduction d'un facteur $\approx d/2$ quand $d \to \infty$. Cette procédure peut être illustrée par une structure en arbre où les valeurs à multiplier se trouvent dans les branches et où chaque noeud équivaut à un appel de la fonction récursive qui accumule les multiplications précédentes et effectue le branchement respectif, comme indiqué dans la figure 1 pour le cas $d = 3$ et où $b_i = P(q_k|x_i)\ \forall i \in J$ pour simplifier la notation. Même si dans cet exemple le gain n'est pas significatif (4 multiplications au lieu de 5), dans le cas de 8 bandes il y aura 247 multiplications au lieu de 769, 65519 contre 458753 avec 16 bandes.

3. Expériences

3.1. *Les reconnaisseurs spectre entier*

Les paramètres d'entrée des ANNs sont du type PLP, avec un prétraitement J-RASTA [HM94] pour la moitié de l'expérience. Les entrées de l'ANN comportent 9 vecteurs acoustiques consécutifs, fournissant une information contextuelle importante au système. Les sorties correspondent aux 27 phonèmes significatifs de la base de données. Les vecteurs acoustiques pour le spectre entier comprennent 12 coefficients PLP (ou J-RASTA-PLP) et l'énergie (ainsi que les dérivées premières et deuxièmes de ces paramètres). Pour les ANNs du spectre entier nous avons choisi 1750 unités cachées. En clair le taux d'erreur au niveau du mot est de 8.0 % pour le système spectre entier sur les J-RASTA-PLPs et de 7.1 % pour le système spectre entier sur les PLPs.

3.2. *Les 4 reconnaisseurs sous-bandes*

Nous avons travaillé avec $d = 4$ sous-bandes. Pour ne pas rajouter des termes de dépendance inter-bandes dans nos modèles FC et AFC, il est précieux de veiller à choisir des bandes spectrales du signal qui se recouvrent au minimum mais dont l'union représente le spectre entier. Dans ce but, nous avons redéfini les sous-bandes qui ne tenaient pas compte de ces contraintes d'indépendance dans des études précédentes [BG99, HMB99]. De plus, nos expériences ont confirmé que la première bande critique n'est pas pertinente en parole téléphonique, nous l'avons donc supprimé. Ainsi nous avons choisi un ensemble homogène de quatre sous-bandes décrit dans la table 1 et qui permet toujours de modéliser un formant par sous-bande. L'ordre des analyses LPC et le nombre de coefficients extraits ont été optimisés sur plusieurs expériences. Dans le cas des fusions de sous-bandes i en une combinaison J, les ordres LPC ainsi que le nombre de coefficients extraits sont la somme de ceux des sous-bandes contenues dans J. Ainsi le nombre de paramètres dans le modèle FC et le modèle AFC sont identiques.

sous-bandes	en Hz	LPC	# coeff.
1	115-629 Hz	3	5
2	565 1370 Hz	3	5
3	1262 2292 Hz	2	3
4	2122 3769 Hz	2	3
134	115-629 Hz, 1262-3769 Hz	7	11

Table 1: Définition des 4 sous-bandes (coupure à 3dB) et des paramètres extraits. Exemple de combinaison : 134, montrant le calcul du nombre de paramètres des flux : somme de ceux des sous bandes, garantissant un nombre de paramètres constant entre FC et AFC. Le faible recouvrement fréquentiel entre sous-bandes est du aux filtres PLP des bandes critiques.

Les systèmes HMM/ANN hybrides pour les sous-bandes correspondent aux systèmes HMM/ANN hybrides spectre entier de base, la seule différence étant le nombre d'entrées, le nombre d'unités cachées restant proportionnel au nombre d'entrée. Les ANNs

des sous-bandes et des combinaisons de sous-bandes ont entre 666 et 1750 unités cachés.

3.3. Base de Données et Bruitage

Base de Données Numbers'95 Nous travaillons sur la base NUMBERS'95 qui contient des chiffres prononcés en continu en anglais, provenant de lignes téléphoniques. Pour l'entraînement de nos ANN nous avons utilisé 3590 phrases, et pour les expériences 200 phrases de l'ensemble de test.

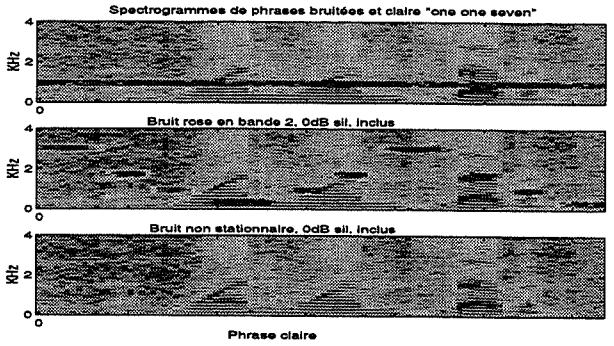


Figure 2: Illustration de bruitage par "bruit coloré" dans la bande 2 (haut), du bruit non stationnaire (milieu), de la parole non-bruitée pour la même phrase «one one seven» (bas).

Définition des bruits idéaux Afin de tester nos modèles dans des conditions idéales nous avons généré une série de quatre "bruits colorés" centrés sur les sous-bandes i ($i = 1..4$) du modèle et n'affectant qu'une sous-bande à la fois (générés par filtres trapézoidaux, 300 Hz de large, fréquence centrale égale à celle des sous-bandes du modèle hors recouvrement). Enfin pour tester l'effet de la robustesse des modèles FC et AFC suivant la distribution du bruit, nous construisons un bruit non stationnaire à partir des mêmes bruits colorés en conservant une répartition homogène du bruit sur les sous-bandes. Des pavés de 125 ms sont régulièrement tirés des sous-bandes 1, 2, 3, 4, 4, 3, 2 et 1 (voir figure 2), comme dans [BGEB98]. Pour clarifier les expériences les niveaux des bruits ajoutés ont été calculés silences inclus et ajoutés phrase par phrase.

3.4. Résultats et discussions

Nous présentons à la figure 3 les résultats des tests sur des bruits colorés dans les différentes sous-bandes 1 à 4, en utilisant les approches FC et AFC et des paramètres PLP. Pour comparaison, cette figure montre également les courbes correspondantes au système spectre entier testé dans les mêmes conditions. Nous constatons que non seulement l'approche FC mais aussi son approximation présentent de très bonnes propriétés de robustesse aux bruits colorés. En effet la reconnaissance peut s'effectuer de façon très fiable sur les composantes non bruitées et la contribution des flux bruités intégrés dans le FC ou AFC est peu perturbante car leur distribution de probabilités *a posteriori* a une forte entropie (distribution plus uniforme). Nous reviendrons sur cette analyse dans le cas de l'AFC.

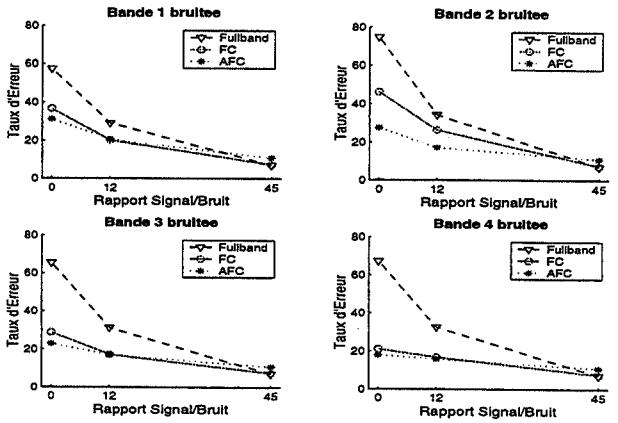


Figure 3: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **PLP**. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée dans des sous-bandes 1 à 4.

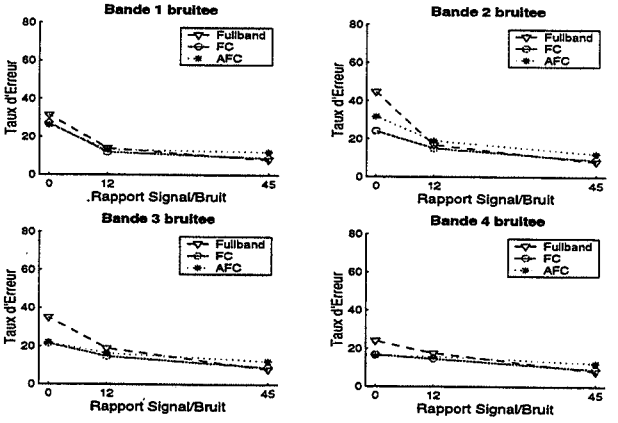


Figure 4: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **J-RASTA**. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée stationnaire dans des sous-bandes 1 à 4.

La figure 4 présente les résultats avec les paramètres caractérisques J-RASTA-PLP. On voit que le filtre J-RASTA est capable de supprimer une partie des interférences dues au bruit coloré ce qui améliore les résultats sur tous les bruits colorés comparés aux résultats avec PLP seul. Mais là encore le FC en tout RSB est plus performant que le système spectre entier (ou égale en claire). Il en est de même pour l'AFC sauf en parole claire où les performances sont moindres car l'expert spectre entier est estimé, la perte des termes de dépendance inter-bande se faisant alors plus ressentir. De même les résultats FC ou AFC obtenus avec les PLPs et J-RASTA-PLPs dans le cas du bruit additif non-stationnaire (fig. 5) indiquent une robustesse plus élevée que celle du spectre entier, même si on note une hausse génèrale des taux d'erreur par rapport aux deux expériences précèdentes (il y a en effet plus de trames bruitées par effet de bord). Dans ces conditions le processus J-RASTA est mis en défaut par rapport au PLP simple, ce qui montre l'inadéquation de J-RASTA à un bruit si non-stationnaire.

Une propriété intéressante de nos modèles est que l'AFC montre une robustesse égale ou supérieure au

FC, sauf dans le cas où la bande 2 est bruitée. Nous avons mesuré que la bande 2 est la plus performante en RAP propre parmi les 4 sous-bandes, l'AFC est donc particulièrement pénalisée dans ce cas. Dans tous les autres cas, les bonnes performances de l'AFC sont dues à une meilleure exploitation de la redondance du signal. En effet en condition bruitée les entropies des vecteurs de probabilités *a posteriori* augmentent autrement dit ces vecteurs ont une distribution plus applatie : les probabilités *a posteriori* tendent vers l'équiprobabilité. Donc dans le modèle AFC qui procède par produits et normalisations, les sous-bandes bruitées affectent peu la distribution des vecteurs porteurs d'information correcte qui eux sont trés discriminants. Le modèle AFC joue donc le rôle d'un filtre : l'information provenant d'une sous-bande de données claires est mieux conservée en sortie du modèle AFC, alors qu'elle est noyée avec les données bruitées dans le cas du modèle FC. En FC dès qu'un flux est partiellement bruité, les probabilités *a posteriori* issues directement de l'expert correspondant sont globalement détériorées et irrécupérables comme le montre [LC97], ce qui défavorise le FC.

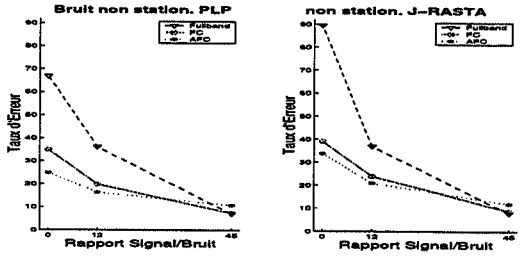


Figure 5: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **PLP** (à gauche) et des **J-RASTA** (à droite). Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée non-stationnaire variant entre les sous-bandes 1 à 4.

4. Conclusions et Perspectives

Nous avons montré qu'en utilisant l'approche sous-bandes FC ou son approximation AFC, même dans le cadre le plus simple de la pondération équiprobable, la robustesse du RAP est plus élevée que celle d'un modèle spectre entier J-RASTA pour les bruits à bande limitée stationnaires et non-stationnaires. Avec des modèles à résolution spectrale supérieure des résultats identiques sont attendus en bruits naturels, ce qui est réalisable efficacement avec la procédure récursive présentée dans ce papier. Une amélioration en parole propre du modèle AFC est accessible en utilisant le vrai estimateur spectre entier, les performances en claires pour AFC et FC sont alors comparables suivant nos récentes expériences. Cette étude comparative entre FC et AFC a mis en évidence que le modèle AFC a des performances sensiblement égales en conditions bruitées, et parfois même supérieures au FC, ce qui a été discuté. Des études publiées ou en cours montrent un accroissement de robustesse des modèles lorsque les poids sont variables selon les performances relatives des flux en condition claire comme les poids « Expectation Maximization » et « Least Mean Square Error » [MHGB00]. Mais les gains en robustesse sont plus grands avec l'usage de poids adaptatifs au bruit. Ces derniers peuvent être basés sur un classique RSB [HMB99] ou suivant une approche «CASA» à partir d'indices d'harmonicité ou de localisation spatiale, poids développés et testés en AFC dans [BG99, GBT99, Glo00].

Remerciements:

Ce travail a été soutenu par les projets Européens TMR SPHEAR et LTR RESPITE, et l'office Fédéral de l'Education et de la Science (OFES).

Bibliographie

[BG99] F. Berthommier and H. Glotin. A new snr-feature mapping for robust multistream speech recognition. In Berkeley University Of California, editor, *Proc. Int. Congress on Phonetic Sciences (ICPhS)*, volume 1 of *XIV*, pages 711–715, San Francisco, August 1999.

[BGEB98] F. Berthommier, H. Glotin, Tessier E., and H. Bourlard. Interfacing of casa and partial recognition based on a multistream technique. In *Int. Conf. on Spoken Language Processing (ICSLP)*, volume 4, pages 1415–1419, 1998.

[Cer99] C. Cerisara. *Contribution de l'approche Multi-Bande à la reconnaissance automatique de la parole.* PhD thesis, Doctorat de l'Institut National Polytechnique de Lorraine, Nancy, France, Sept. 1999.

[Dup00] S. Dupont. *Études et Développement de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole.* PhD thesis, Doctorat de l'Université de Mons, Belgique, 2000.

[GBT99] H. Glotin, F. Berthommier, and E. Tessier. A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, volume 5, pages 2351–2354, september 1999.

[Glo00] H. Glotin. *Elaboration et étude comparative d'un système adaptatif de reconnaissance robuste de la parole en sous-bandes : incorporation d'indices primitifs F0 et ITD.* PhD thesis, Doctorat de l'Institut National Polytechnique de Grenoble, Avril 2000.

[HM94] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[HMB99] A. Hagen, A. Morris, and H. Bourlard. Different weighting schemes in the full combination subbands approach in noise robust asr. In *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999.

[HTP96] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. pages 462–465, 1996.

[LC97] R. Lippmann and B. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In ESCA, editor, *Eurospeech'97*, pages 37–40, 1997.

[MHGB00] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust asr. *Speech Communication*, to appear, 2000.

La combinaison de la transformation linéaire et du MAP pour l'adaptation de modèles de langage[1]

D.Janiszek, F.Béchet, R. De Mori

LIA - CERI
Université d'Avignon, BP 1228
84911, Avignon Cedex 9, France
Tél.: ++33 (0)490 84 35 00 - Fax: ++33 (0)490 84 35 01
{ david.janiszek, frederic.bechet, renato.demori} @lia.univ-avignon.fr

ABSTRACT

The paper discusses the integration of various language model adaptations. Ways of integrating Maximum A Posteriori (MAP) adaptation and linear transformation of bigram probability vectors are introduced and evaluated. For adaptation corpora of less than 15000 words, it is shown that such an integration leads to a substancial reduction of perplexity with respect to direct LM probability estimation with back-off from the adaptation data. A reduction of the word-error rate is also observed for the corpora of less than 5000 words.

1. INTRODUCTION

La plupart des systèmes de reconnaissance automatique de la parole (SRAP) existants établissent des hypothèses de mots en estimant la probabilité d'une séquence de mots par le produit des probabilités conditionnelles des mots sachant leur historique (1). Ainsi la probabilité d'une séquence de N mots W_1^N est :

$$P(W_1^N) = \prod_{n=1}^{N} P(w_n \mid H_n) \qquad (1)$$

Où w_n est le n-ième mot de la séquence et H_n son historique. Toutefois l'estimation de toutes les probabilités conditionnelles est irréalisable, d'où l'utilisation de classes d'historiques. Les modèles de langage (ML) fournissent des distributions de probabilités $P(w_i|h_i)$ pour chaque mot w_i du vocabulaire et pour chaque classe d'historique h_i, y compris la classe nulle.

Les probabilités $P(w_i|h_i)$ sont estimées à partir d'un corpus d'entraînement. En pratique leurs valeurs et leurs précisions dépendent du corpus, de son adéquation au domaine de l'application et de sa taille. En effet un ML estimé sur un domaine D1 donne habituellement de médiocres résultats, comparé à un ML estimé sur un domaine D2, lors de leur utilisation dans ce même domaine D2. D'autre part, on remarque qu'en deçà d'une certaine taille, le corpus ne contient pas suffisamment de données pour estimer correctement un ML.

En pratique, lors du développement d'une nouvelle application dans un domaine spécifique D2, il arrive que le corpus disponible soit de taille insuffisante. On cherche alors à exploiter un ML bien estimé issu d'un autre domaine D1. Afin d'obtenir un ensemble de données suffisant et approprié au domaine D2, on adapte le ML estimé sur le domaine D1 aux observations (données d'adaptation) issues du domaine D2. Plusieurs méthodes d'adaptation sont disponibles :

- Interpolation linéaire entre un domaine général et le domaine spécifique [Sey97]
- Interpolation log-linéaire [Kla98]
- Replis du modèle spécifique vers un modèle général [Bes95]
- Récupération de documents pertinents dans le domaine D2 [Iye99]
- Discrimination minimale et maximum d'entropie [Che98]
- Maximum a posteriori sur les probabilités [Fed99]
- Transformation linéaire des vecteurs de comptes de bigrammes dans un espace réduit [Jan99]

Mais la dispersion des données augmente avec la taille du vocabulaire, et des événements linguistiques peuvent être absents d'un corpus, même de grande taille. Pour cette raison, il est souhaitable de concevoir un algorithme estimant aussi les probabilités d'événements absents du corpus d'adaptation. L'objectif de l'adaptation est d'améliorer le taux de reconnaissance du SRAP de la nouvelle application, par rapport à la seule utilisation des données receuillies dans son domaine.

[1] Cette recherche est soutenue par le *CNET – France Télécom* sous le contrat 971B427

2. LA TRANSFORMATION LINÉAIRE DANS UN ESPACE RÉDUIT

Un critère largement utilisé pour comparer des MLs est la mesure de leur perplexité ; on considère la perplexité la plus basse comme étant la meilleure. Mais il n'y a pas de preuve théorique que l'utilisation du ML de perplexité inférieure, dans un SRAP, implique un meilleur résultat de reconnaissance. Néanmoins des preuves empiriques montrent qu'une baisse substantielle de perplexité améliore souvent les résultats de reconnaissance.

Une nouvelle approche est proposée pour adapter un ML général, bien estimé, en ML approprié à une nouvelle application. Cette adaptation est basée sur une transformation linéaire des événements linguistiques. Le critère d'optimisation de cette transformation est la minimisation de la perplexité du ML adapté mesurée sur une partie du corpus d'observation.

Soit $P = \{p_{ij}\}$ une matrice de dimensions IxJ dont l'élément générique $\{p_{ij}\}$ représente la fréquence d'observation du mot w_i avec le contexte historique h_j. Soit W_i^P le i-ème vecteur de la matrice P, ses J éléments sont les probabilités du mot w_i avec tous les historiques possibles. On cherche alors pour chaque vecteur une matrice A_i telle que le vecteur de probabilités adaptées s'obtienne par le produit suivant :

$$W_i^a = A_i . W_i^P \qquad (2)$$

L'élément générique $\{a_{ijk}\}$ de la matrice A_i peut être estimé par un processus d'optimisation tendant vers une diminution de la perplexité du ML et satisfaisant aux contraintes suivantes :

- les éléments des vecteurs sont positifs

$$\sum_{k=1}^{J} a_{ijk} p_{ij} \geq 0 \qquad (3)$$

- la somme des probabilités de chaque historique doit être égale à un.

$$\sum_{j=1}^{J} \sum_{k=1}^{J} a_{ijk} p_{ij} = 1 \qquad (4)$$

En utilisant le ML dont les probabilités sont obtenues par cette transformation linéaire, la perplexité peut être exprimée de la manière suivante :

$$PP = 2^{-\frac{1}{N_a} \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} \log \sum_{k=1}^{J} a_{ijk} p_{ij}} \qquad (5)$$

Où N_a est le nombre de mots du corpus, et n_{ij} le nombre d'occurrences du mot w_i avec l'historique h_j .

Dans [Jan99] il est montré que la minimisation de la perplexité, comme critère d'optimisation, conduit a une adaptation dont les probabilités calculées se rapprochent des probabilités des événements observés dans le corpus d'adaptation. Ainsi, lorsque le corpus d'adaptation est limité, nombre d'événements linguistiques ne sont pas observés, et les probabilités correspondantes ne sont pas adaptées. Mais en pratique, pour surmonter cette difficulté, on calcule cette transformation après la projection des vecteurs dans un espace réduit.

Cet espace est obtenu par la décomposition en valeurs singulières (SVD) d'une matrice représentant un corpus général. qui est une représentation typique de la langue et donc de ses événements linguistiques. L'utilisation d'un espace réduit permet une adaptation des informations qui contribuent le plus à la structure du domaine, et une approximation des autres. Ainsi on peut estimer les probabilités d'événements, même s'ils n'apparaissent pas dans le corpus d'adaptation, lors du calcul du ML adapté.

Dans [Jan99] une méthode est proposée pour estimer une transformation affine A* qui permet pour une ou plusieurs classes d'adapter les vecteurs W_i^G d'un corpus général vers les vecteurs W_i^P du corpus d'adaptation. Cette transformation est réalisée sur les comptes des bigrammes, ce qui est équivalent à l'utilisation des probabilités (en dehors de l'utilisation de facteurs de replis), mais qui permet un respect plus aisé des contraintes (3) et (4). Cette méthode entraîne des baisses significatives de perplexité sur des données appartenant au corpus de dialogue *AGS* [Sad96] fourni par le *CNET – France Télécom*, lorsque la taille du corpus d'apprentissage est inférieure à 15000 mots.

3. COMBINAISON DE LA TRANSFORMATION LINÉAIRE ET DU MAP

Dans [Fed96] il est montré que l'adaptation de MLs par un MAP peut être considéré comme une interpolation linéaire des fréquences relatives. Et dans [Fed99] il est montré que l'adaptation d'un ML d'unigrammes peut être réalisée par une interpolation linéaire entre les probabilités à priori d'un ML général et celles obtenues avec le corpus d'adaptation. On utilise donc la formule suivante :

$$c_{MAP} = \lambda c_A + (1-\lambda) c_B \qquad (6)$$

Avec :

$$\lambda = \frac{m_A}{m_A + m_B} \qquad (7)$$

Où c_A et c_B sont respectivement les comptes adaptés par la transformation linéaire, et les comptes du corpus d'adaptation. Et m_A et m_B sont respectivement la masse des comptes adaptés et la masse des comptes du corpus d'adaptation. Evidemment cela n'est possible que lorsque les masses de chaque ensemble ont le même ordre de grandeur. [Fed96] propose d'ailleurs l'introduction d'un paramètre pour atténuer la

différence relative entre les masses, mais cela reste une interpolation linéaire.

Ainsi le processus d'adaptation des données se décompose en trois phases :

- Projection des données dans un espace réduit.
- Adaptation par une transformation linéaire
- Ajustement des données adaptées par un MAP, entre les données adaptées et les données observées dans la nouvelle application (ou données d'adaptation).

4. EXPÉRIENCES ET RÉSULTATS

Les expériences d'adaptation ont été réalisées en utilisant un corpus spécifique (70000 mots) établit à partir de phrases de dialogue collectées par le système *AGS* du *CNET – France Télécom* et un corpus général d'articles de presse du journal *Le Monde* (40 millions de mots)

Le vocabulaire est composé des mots que l'on souhaite reconnaître, dans notre cas il s'agit des 823 mots de notre corpus d'entraînement issu de phrases collectées par le système *AGS*, et de 102 classes syntaxiques (POS) utilisées par le tagueur du LIA.

La matrice du corpus général a été obtenue à partir du corpus du *Monde*, en gardant uniquement les mots du vocabulaire d'*AGS* (7 millions de bigrammes). Et pour les classes l'intégralité du corpus est utilisée, chaque mot ayant été étiqueté.

L'espace réduit est obtenue par une SVD de la matrice du corpus général, et la sélection des q valeurs singulières s_i, ordonnées de manière décroissante et telles que $s_1/s_q \approx 10^3$. La base de l'espace réduit correspond donc aux q vecteurs correspondant. (En pratique cela donne : q=43 pour les classes et q=412 pour les mots).

Les MLs sont calculés avec facteurs de replis, selon la méthode de Good-Turing, par le CMU-SLM Toolkit [Ros94]. Lors des expériences de reconnaissance le facteur de combinaison score acoustique/score linguistique est calculé sur les données brutes.

L'expérience mise en place vise à déterminer les tailles de corpus pour lesquelles la combinaison de la transformation linéaire et du MAP apportent un gain significatif. Ainsi, dans un premier temps, on mesure la perplexité des MLs sur un corpus de test de 400 phrases, soit environ 2000 mots.

La figure 1 représente l'évolution de la perplexité des bigrammes de classes syntaxiques (POS), pour différentes techniques, en fonction de la taille du corpus. La taille correspond aux n premiers éléments du corpus d'apprentissage étiqueté en classes. A partir de cet extrait on calcule un ML pour chaque technique :

- Le ML 'Données Brutes' correspond à l'utilisation des données sans autre traitement.
- Le ML 'LT' correspond à l'utilisation des données adaptées par transformation linéaire avec une seule classe puis discrétisation des valeurs.
- Le ML 'MAP' correspond à l'utilisation des formules (6) et (7)
- Le ML 'LT+MAP' correspond à la combinaison de la transformation linéaire et du MAP telle que décrite à la section précédente.

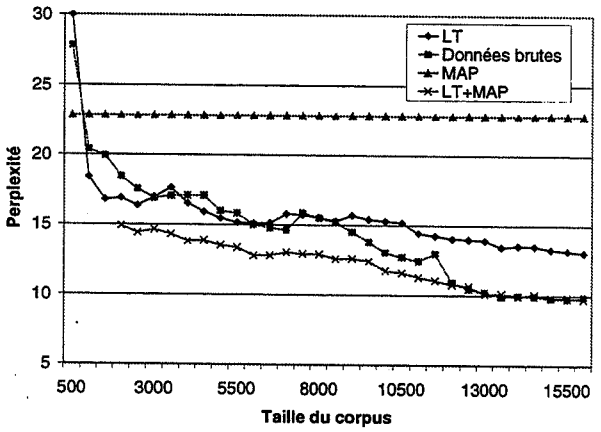


Figure1: Comparaison des techniques d'adaptation, par une mesure de perplexité en fonction de la taille d'un corpus de classes POS.

De la même manière, la figure 2 montre l'évolution de la perplexité des bigrammes de mots, pour les différentes techniques, en fonction de la taille du corpus. Les modèles de langages sont calculés de façon similaire mais avec les mots.

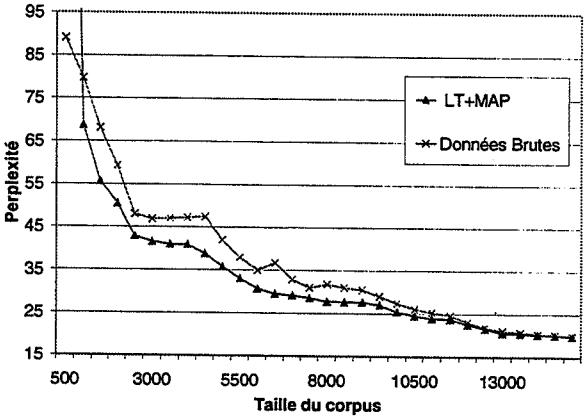


Figure2: Comparaison des techniques d'adaptation, par une mesure de perplexité en fonction de la taille d'un corpus de mots.

La figure 1 montre qu'en terme de perplexité, le MAP améliore les résultats de la transformation linéaire. La combinaison des deux techniques entraîne une

amélioration sensible des résultats par rapport à l'utilisation des données brutes pour des corpora de classes inférieurs à 15000 mots. Au delà l'apprentissage sur le corpus se révèle plus efficace. Le gain maximal en perplexité passe d'environ 16% à 19%.

La figure 2 vérifie cette amélioration avec l'utilisation des mots : le gain maximal est de 18 %, et la combinaison est intéressante pour des corpora inférieurs à 15000 mots. Au delà les données d'apprentissage sont suffisantes.

Table 1: Taux d'erreurs de reconnaissance en fonction de la taille du corpus et des données utilisées.

Taille du corpus (en mots)	Données brutes	Données adaptées
4000	44,24 %	40,44 %
8000	38,51 %	38,75 %
16000	36,93 %	37,27 %
32000	36,18 %	37,07 %

Les tests de reconnaissance réalisés sur 393 treillis recueillis par le système *AGS* montrent une amélioration des performances plus modérée. En effet les taux de reconnaissance des MLs bruts et adaptés (LT+MAP) sont proches quelle que soit la taille du corpus d'apprentissage. Toutefois l'adaptation ne se révèle effectivement utile que pour des corpora de taille réduite (ici moins de 5000 mots). Au delà, aucune amélioration n'est observée.

5. Conclusions

La combinaison des deux techniques d'adaptation apporte des baisses significatives de perplexité, mais l'amélioration des taux de reconnaissance est plus modérée. Les baisses de perplexité semblent donc insuffisantes. Cependant, il faut remarquer que seuls 226 treillis contiennent leur phrase de référence, ce qui limite le gain potentiel.

Néanmoins les résultats pourraient être améliorés par :

- le calcul d'un meilleur coefficient d'interpolation ; en effet le coefficient idéal permet une baisse plus importante du taux d'erreurs de reconnaissance.
- la récupération des ruptures de séquence ; en effet près de 16 % des phrases comportent une omission de mot. Cette rupture de séquence ne peut pas être correctement récupérée par un bigramme.
- l'utilisation d'un corpus général mieux adapté ; en effet *Le Monde* est un corpus d'articles de presse. Il est éloigné, en terme d'événements linguistiques, d'un corpus issu d'une application de dialogue. On observe d'ailleurs une amélioration de la perplexité pour des corpora inférieurs à 35000 mots lorsque l'espace réduit est calculé à partir de la totalité du corpus spécifique. Certaines données sont alors utilisées deux fois . Pour vérifier cette tendance il faudrait utiliser un grand corpus de dialogue général ; cependant on se heurte à un problème de disponibilité.

Bibliographie

[Bes95] Besling S., Meier H.G. (1995) "Language Model Speaker Adaptation", Eurospeech 95, pp. 1755-1758

[Che98] Chen S.F., Seymore K., Rosenfeld R. (1998) "Topic Adaptation For Language Modeling Using Unnormalized Exponential Models", IEEE Intl. Conf. On Acoustics, Speech ans Signal Processing, Seattle, USA

[Fed96] Federico M. (1996) "Bayesian Estimation Methods For N-Gram Language Model Adaptation", ICSLP 96, Vol. 1, pp. 240-243

[Fed99] Federico M.,De Mori R. (1999) "Language Model Adaptation", Computational Models Of Speech Pattern, K. Ponting ed. Springer-Verlag, Berlin, New-York

[Iye99] Iyer R., Ostendorf M. (1999) "Modeling Long Distance Dependence In Language : Topic Mixtures vs. Dynamic Cache Models", IEEE Transactions on Speech and Audio processing, SAP-7(1) , pp30-39

[Jan99] Janiszek D., De Mori R., Béchet F., Matrouf D., Mokbel C. (1999) "New Language Model Adaptation Algorithm Based On The Definition Of Cardinal Distance ", ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, Kloster Irsee, Allemagne, pp.97-100

[Kla98] Klakow D. (1998) "Log-Linear Interpolation Of Language Models", Proc. International Conference on Spoken Language Processing, Sydney, Australie

[Ros94] Rosenfeld R .(1994) "Adaptive Statistical Language Modeling : a Maximum entropy approach ", PhD Thesis Carnegie Mellon University , Pittsburgh

[Sad96] Sadek D., Ferrieux A.,Cozannet A., Bretier P., Panaget F., Simoni J. (1996) "Effective Human-Computer Cooperative Spoken Dialogue : The AGS Demonstrator", ICSLP, Vol. 24, pp. 456-465.

[Sey97] Seymore K., Rosenfeld R. (1997), "Using Story Topics For Language Model Adaptation", Eurospeech, pp. 1987-1990.

Système hybride markovien/K-plus proches voisins pour la reconnaissance de la parole continue

Fabrice Lefèvre[1,2], *Claude Montacié*[1] *et Marie-José Caraty*[1]

1-Laboratoire d'Informatique de Paris 6 - 4, place Jussieu – 75252 Paris Cedex 5
2-Laboratoire d'Informatique d'Avignon - BP 1228– 84911 Avignon Cedex 09, France
Tél.: +33 (0)4 90 84 35 32 - Fax: +33 (0)4 09 84 35 01
Mél: Fabrice.Lefevre@lia.univ-avignon.fr - http://www.lia.univ-avignon.fr

ABSTRACT

In this paper, a new hybrid speech recognition system is presented : the K-nearest neighbours HMM-based system. The basis idea is to compute the HMM states emission probabilities using the K-nn estimator instead of the widespread estimator based on mixtures of gaussian functions.

First, the performances of the K-nn estimator are evaluated on speech data by a frame identification. Then, an optimal protocol is derived for the development of a basis K-nn/HMM system. This system is evaluated through segmental recognitions. Afterwards, two methods are proposed to upgrade its performances.

1. INTRODUCTION

Les modèles de Markov cachés (*Hidden Markov Model,* HMM) sont à l'origine de la majorité des avancées récentes en reconnaissance de la parole (RAP) continue. Ces modèles gèrent les distorsions temporelles du signal de parole en s'appuyant sur des densités de probabilité pour modéliser les distorsions en fréquence. Ces densités de probabilités sont communément estimées par des sommes pondérées de gaussiennes (*Gaussian Mixture Model,* GMM). Nous proposons de les remplacer par un estimateur de densité de probabilité plus efficace : l'estimateur des K plus proches voisins (K-ppv). La règle de décision déduite de cet estimateur présente l'intérêt d'avoir une erreur asymptotique de classification faible (proche de l'erreur optimale de Bayes) et qui diminue avec l'augmentation de K. De plus, il est discriminant par construction et ne nécessite pas d'apprentissage.

Dans ce papier, un panorama des travaux déjà réalisés est proposé, complété par une présentation des derniers développements [1]. Dans la section 2, l'estimateur des K-ppv est évalué comme opérateur de reconnaissance statique de spectres à court-terme de parole. Ses performances se révèlent supérieures à celles de l'estimateur de l'état de l'art à base de GMM. Les adaptations nécessaires à son intégration dans un système de reconnaissance markovien sont ensuite développées dans la section 3. Puis le système HMM/K-ppv est évalué sur la base de données TIMIT sur des tâches de reconnaissance phonétiques en comparaison avec un système HMM/GMM.

Finalement, dans la section 4, deux approches sont envisagées pour faire évoluer le système de base HMM/K-ppv vers un système de référence : l'introduction de coefficients delta dans les paramètres d'entrées du système et la modélisation contextuelle.

2. ESTIMATEUR DES K-PPV

Avant son introduction dans le formalisme markovien, l'estimateur des K-ppv est évalué localement afin de valider ses propriétés théoriques sur des données de parole.

2.1 Validation en identification locale

Le principe de l'identification locale (ou trame-à-trame) est d'assigner chaque trame de l'ensemble d'évaluation à une classe phonétique sans considérations segmentales. La règle de décision du maximum *a posteriori*, déduite de l'estimateur des K-ppv, sera utilisée : la classe reconnue est la classe associée au plus grand nombre de ppv de la trame considérée.

Les expériences sont menées sur la base de données TIMIT. L'ensemble d'apprentissage est composé de 1.124.823 trames (3696 phrases) et le test de 57.919 trames (192 phrases). Calculée toute les centi-secondes sur une fenêtre d'analyse de 25ms, une trame est représentée par 12 MFCC et l'énergie à court-terme. Pour chaque trame de l'ensemble d'apprentissage et de test, l'estimateur K-ppv est obtenu à partir de ses K plus proches voisins (au sens d'une distance de Mahalanobis) dans l'ensemble d'apprentissage. 39 classes phonétiques ont été considérées [2]. Pour cette expérience, nous avons fait varier K de 1 à 50. Les résultats confirment la diminution du taux d'erreur avec l'augmentation de K (cf. figure 1).

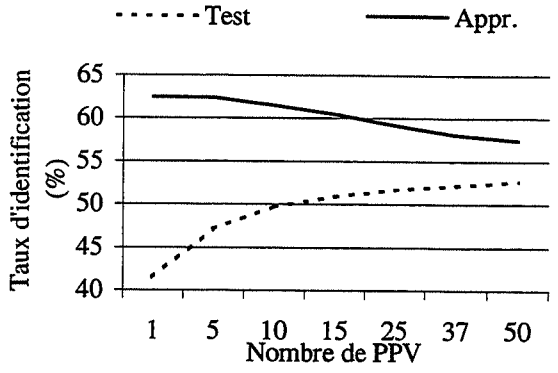


Figure 1. Evolution du taux d'identification locale sur TIMIT de la règle des K-ppv en fonction de K.

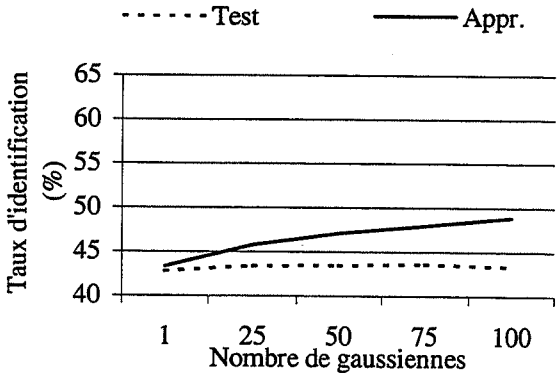


Figure 2. Evolution du taux d'identification locale sur TIMIT des GMM en fonction du nombre de gaussiennes par mixture

2.2 Comparaison GMM/K-ppv

L'expérience précédente a été reprise avec l'estimateur à base de GMM. Un GMM est initialisé pour chaque classe phonétique par l'algorithme des k-moyennes puis ses paramètres sont affinés à l'aide de l'algorithme de Baum-Welch. Pour le GMM, 48 classes phonétiques sont apprises, regroupées en 39 lors du calcul du score.

Des GMM de 1 à 100 gaussiennes par mélange ont été considérés. On constate que l'augmentation du nombre de gaussiennes n'améliore les résultats de test que faiblement avec un maximum pour 75 gaussiennes (cf. figure 2). Les taux d'identification locale des estimateurs K-ppv et GMM sont comparés (voir figure 3).

Pour chaque ensemble d'évaluation, les deux meilleures configurations ont été considérées. L'estimateur des 50-ppv obtient des résultats supérieurs à l'estimateur GMM de l'ordre de 9% sur le test et jusqu'à 14% en auto-cohérence.

3. DEVELOPPEMENT D'UN SYSTEME HMM/K-PPV DE BASE

Les adaptations nécessaires à l'introduction de l'estimateur des K-ppv dans les HMM sont présentées. Puis, un système HMM/K-ppv de base est évalué par des expériences de reconnaissance phonétique.

3.1 Introduction de l'estimateur des K-ppv dans le formalisme markovien

La re-formulation de l'expression des probabilités d'émission des états par l'estimateur des K-ppv pose le problème de l'association des trames aux états markoviens.

Par le biais d'un étiquetage de référence, une trame est associé au HMM représentant sa classe phonétique. Mais ensuite, l'association de la trame à un des états du HMM doit être obtenue automatiquement dès lors qu'aucune expertise n'est disponible à ce niveau. Par ailleurs, une assignation univoque paraît une solution peu souhaitable dans la mesure où elle est obtenue automatiquement.

Pour résoudre ce problème, nous avons eu recours à la notion de degré d'appartenance empruntée à la théorie des ensembles flous. A chaque observation o_t de l'ensemble d'apprentissage est attribué un coefficient d'appartenance $u_i(o_t)$ sur chaque état i qui vérifie :

$$u_i(o_t) \in [0,1] \qquad \sum_{i=1}^{S} u_i(o_t) = 1$$

avec S le nombre total d'états dans le système. L'expression de la probabilité d'émission sur l'état i pour l'observation o_t peut alors s'exprimer :

$$b_i(o_t) = \frac{\sum_{k=1}^{K} u_i(ppv(k,o_t))}{U_i} \quad \text{avec} \quad U_i = \sum_{j=1}^{O} u_i(o_j)$$

où $ppv(k,o_t)$ est une fonction qui renvoie le $k^{ième}$ ppv de l'observation o_t et U_i est la somme des coefficients d'appartenance à l'état i sur les O observations de l'ensemble d'apprentissage.

Les coefficients d'appartenance deviennent les paramètres du système HMM/K-ppv. Ils sont appris par l'algorithme de Baum-Welch. Pour cela, une formule de réestimation a été obtenue dont nous avons montré la convergence selon le critère du maximum de vraisemblance [3].

3.3 Evaluation comparée des systèmes HMM/GMM et HMM/K-ppv

Les résultats de l'identification segmentale et du décodage acoustico-phonétique (DAP) des systèmes HMM/50-ppv et HMM/32-GMM sur TIMIT sont mis en relation dans la figure 3. Les segments de silence sont préalablement ôtés du signal. Les modèles utilisés sont des monophones à 3 états de type gauche-droit (autorisant un saut de l'état central et un bouclage à l'état courant). Le DAP intègre un bigramme

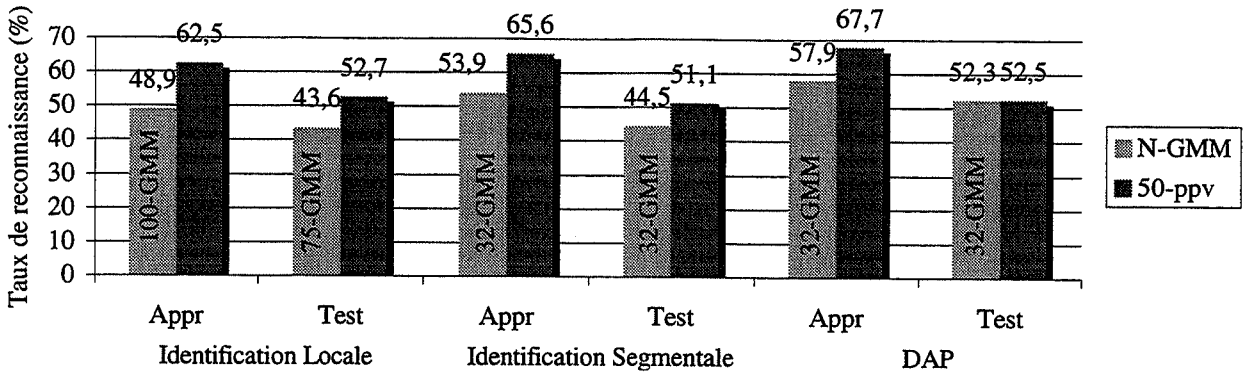


Figure 3. Résultats des systèmes HMM/50-ppv et HMM/N-GMM pour les trois séries d'expériences (identification locale, identification segmentale et DAP) sur l'ensemble d'apprentissage et le core-test de TIMIT.

phonétique avec un facteur d'échelle optimisé sur l'ensemble d'apprentissage.

On observe en reconnaissance segmentale l'écart de performance relevé en identification locale entre les estimateurs K-ppv et GMM. Une écart de l'ordre de 10% est constatée en auto-cohérence pour les trois expériences. Cet écart se réduit à 6,6% sur le test de l'identification segmentale. L'exception notable concerne le test du DAP pour lequel les deux systèmes présentent des taux comparables. Ce résultat est surprenant et devra d'être analysé afin de déterminer pourquoi l'écart de performance n'est pas conservé. Toutefois, il nous a semblé préférable, dans un premier temps, de chercher à d'élever les performances du système HMM/K-ppv au niveau de l'état de l'art.

4. EVOLUTION VERS UN SYSTEME DE REFERENCE

L'évolution des performances du système HMM/K-ppv est envisagée à travers deux techniques de l'état de l'art des systèmes HMM/GMM : les coefficients delta et la modélisation contextuelle.

4.1 Coefficients delta

Une meilleure prise en compte de la dynamique temporelle du signal lors de la reconnaissance est introduite dans le système HMM/K-ppv. Parmi les différentes approches possibles, nous avons retenu l'introduction de paramètres temporels dérivés (coefficients delta) dans l'espace de représentation.

La prise en compte des 2 types de coefficients (statiques et delta) est faite au travers de flux associés à chaque sous-vecteur. La probabilité d'émission d'un état i d'un HMM pour l'observation o_t est alors reconstruite par le produit des probabilités d'émission associées à chacun des flux :

$$b_i(o_t) = \left(b_i^s(o_t)\right)^{\gamma_S} \times \left(b_i^D(o_t)\right)^{\gamma_D}$$

Les valeurs optimales (1,1) des pondérations γ ont été obtenues sur une identification locale. Chaque trame de l'ensemble d'apprentissage possède alors deux ensembles de coefficients d'appartenance aux états : un pour le sous-vecteur des coefficients statiques $\{u^s_i(v)\}$ et l'autre pour celui des coefficients delta $\{u^D_i(v)\}$. Les formules de réestimation pour chaque groupe de coefficients ont été adaptées pour permettre la prise en compte de la participation de l'autre dans le calcul de la probabilité d'émission.

L'introduction des coefficients delta dans les paramètres du système HMM/50-ppv n'a pas permis une amélioration de ses performances en DAP (cf. figure 4). Ce résultat est surprenant dans la mesure où des gains ont néanmoins été observés en identification locale (+3,2%) et segmentale (+2,8%).

Une étude complémentaire visant à mettre en évidence l'influence réelle des coefficients delta dans le système HMM/GMM a été menée [4]. Elle ne nous a pas permis de recréer un phénomène comparable dans le système HMM/K-ppv.

4.2 Modélisation contextuelle

La modélisation contextuelle est appliquée au système HMM/K-ppv. En première approche, une modélisation par biphones droits sera utilisée. Il a été montré que, dans le cadre de TIMIT, l'écart de performance est faible entre les modélisations par triphone et par biphone [5].

Les biphones droits possédant plus de 50 occurrences dans l'ensemble d'apprentissage sont retenus. A partir des 39 classes phonétiques initiales, on obtient 519 modèles (480 biphones droits et 39 monophones afin de modéliser les biphones non représentés) sur les 1521 possibles. Ces modèles sont entraînés de manière identique aux monophones.

Puis, afin d'augmenter la qualité de la modélisation, le partage de paramètres au niveau des états est introduit. La divergence de Kullback-Liebler [6] est utilisée dans sa version symétrique comme mesure de similarité entre les distributions des états. Elle s'applique sur les distributions b_i et b_j des états i et j comme :

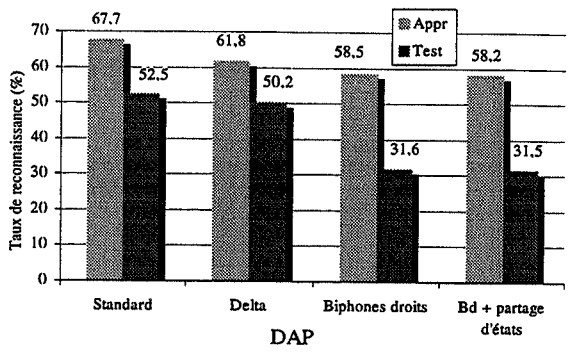


Figure 4. Résultats de DAP sur TIMIT du système HMM/50-ppv dans sa version standard, avec coefficients delta, avec une modélisation par biphones droits simple puis avec un partage de paramètres au niveau des états.

$$KL(b_i \| b_j) = \int_x b_i(x) \log \frac{b_i(x)}{b_j(x)} dx + \int_x b_j(x) \log \frac{b_j(x)}{b_i(x)} dx$$

L'intégrale sur l'espace de représentation est rapportée à une sommation sur toutes les observations o de l'ensemble d'apprentissage :

$$\begin{aligned} KL(b_i \| b_j) &= \sum_o \left\{ b_i(o) \log \frac{b_i(o)}{b_j(o)} - b_j(o) \log \frac{b_j(o)}{b_i(o)} \right\} \\ &= \sum_o \left(b_i(o) - b_j(o) \right) \times \left(\log b_i(o) - \log b_j(o) \right) \end{aligned}$$

Cette procédure a un coût de calcul global en $o(OS^2)$ avec O le cardinal de l'ensemble d'apprentissage et S le nombre total d'états dans le système. Avec O de l'ordre de 1 million et S autour de quelques milliers, cette procédure est très coûteuse. Pour amoindrir son coût, nous avons diminué le nombre de trames utilisées pour l'estimation de la divergence. Ainsi, une population de $O/100$ trames, dont la répartition par classe phonétique est proportionnelle à celle de l'ensemble complet, a été utilisée pour l'estimation ; soit 9.050 trames dans le cas de l'ensemble d'apprentissage de TIMIT (sans silences).

La modélisation par biphones droits conduit à une forte dégradation des résultats de DAP (cf. figure 4). Le réduction du nombre d'états indépendants (de 1557 à 1357) a peu d'effet sur le score. La chute de performance est principalement due à une forte augmentation du taux d'insertion du système.

On peut avancer une hypothèse à l'inadéquation de la modélisation contextuelle au système HMM/K-ppv. Dans le cas des GMM, elle favorise une meilleure représentation par les états markoviens des modes statistiques des données grâce à un découpage préalable en zones plus homogènes. Dans le cas de l'estimateur des K-ppv, elle conduit à introduire une plus grande confusion dans les données par l'augmentation du nombre de classes considérées.

5. Conclusions et Perspectives

Cette étude a montré la faisabilité de l'approche par K-ppv à la RAP continue. Elle a conduit au développement du premier système de reconnaissance hybride HMM/K-ppv. Ses performances en configuration de base sont encourageantes. Elles sont soit meilleures soit équivalentes à celles d'un système HMM/GMM de caractéristiques équivalentes.

L'évolution vers un système de référence par le biais des techniques performantes issues de l'état de l'art des systèmes HMM/GMM n'a pas aboutie. De cette constatation découle la nécessité du développement de techniques spécifiques au système HMM/K-ppv.

Notamment, nous pensons que si la qualité de la représentation locale des données est améliorée avec l'estimateur des K-ppv, elle reste contrainte au niveau segmental par un cadre rigide (modèle gauche-droit à 3 états) qui lui est peu adapté. L'approche par fénones [7] va être étudiée comme méthode d'adaptation de la structure temporelle des HMM à l'estimateur des K-ppv.

Bibliographie

[1] Lefèvre, F., *Estimation de probabilité non-paramétrique pour la reconnaissance markovienne de la parole.* Thèse de l'Université Pierre et Marie Curie, Paris, 2000.

[2] Lee, K.-F., *Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition.* IEEE Transactions on PAMI, **38**(4): p. 599-609, 1990.

[3] Lefèvre, F., C. Montacié, and M.-J. Caraty. *A MLE Algorithm for the K-nn/HMM System. Eurospeech*, Budapest, 1999.

[4] Lefèvre, F., C. Montacié, and M.-J. Caraty. *On the Influence of the Delta Coefficients on a HMM-based Recognition System. ICSLP/SST*, Sydney, 1998.

[5] Barras, C., *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés.* Thèse de l'Université Pierre et Marie Curie, Paris, 1996.

[6] Kullback, S. and R. Leibler, *On Information and Sufficiency.* Ann. Math. Stat., **22**: p. 79-86, 1951.

[7] Bahl, L., *et al.*, *A Method for the Construction of Acoustic Markov Models for Words.* IEEE Transactions on Speech and Audio Processing, **1**(4): p. 443-452, 1993.

Stratégies pour un système de dialogue oral homme-machine

Sophie Rosset, Samir Bennacef †, Lori Lamel
Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133 91403 Orsay Cédex, France
†VECSYS, 3 r. de la Terre de Feu - Les Ulis, 91952 Courtabœuf, France
{rosset,lamel}@limsi.fr sbennacef@vecsys.fr

ABSTRACT

In this contribution we report on our methodology for designing and testing different strategies for dialog management, drawing upon our experience with several travel information tasks. In the LIMSI ARISE system for train travel information we have shifted to a 2-level mixed-initiative dialog strategy improving general navigation and error recovering abilities. The revised dialog strategy has resulted in a 5-10% increase (absolute) in dialog success depending upon the word error rate.

1. INTRODUCTION

Obtenir un taux élevé de succès de dialogue en un miminum de temps (i.e. échanges) reste l'un des objectifs les plus importants dans le développement d'un système de dialogue oral homme-machine. Un dialogue bien dirigé par le système permet d'arriver à des taux de succès élevés, mais généralement au prix d'un grand nombre d'échanges système-utilisateur. Un dialogue complètement libre, où l'utilisateur demande ce qu'il veut quand il veut, peut aboutir en très peu d'échanges, mais en cas de problèmes de reconnaissance ou de compréhension le dialogue risque d'échouer complètement par manque de prédiction. Différentes études [Os99, Hua99] ont montré les avantages d'un système à initiative partagée, où l'utilisateur a majoritairement l'initiative du dialogue sauf dans certains cas bien précis comme par exemple des hésitations trop longues de l'utilisateur ou des erreurs repérées par le système. L'initiative partagée permet ainsi une bonne robustesse face aux erreurs, augmente l'importance des demandes de confirmation [Lav99] ainsi que la capacité à gérer les négations et la référence. Les points importants sur lesquels nous avons concentré nos efforts sont donc la navigation et l'initiative de dialogue (i.e. qui de l'utilisateur ou du système doit diriger le dialogue) et la robustesse du système face aux erreurs. Un autre point important concerne la structuration du domaine (le modèle de la tâche) [Ehr99], dont dépend, au moins partiellement, la capacité de navigation du système, i.e. la capacité de délivrer les informations dans un ordre choisi par l'utilisateur et non préétabli par le système.

Le travail décrit ici a été effectué dans le cadre du projet LE-3 ARISE concernant la demande d'informations par téléphone sur les transports SNCF entre plus de 600 gares en France. Les informations à fournir concernent les horaires, les tarifs, les réductions, les réservations et autres prestations. Notre objectif est de permettre un dialogue avec une structure très ouverte, laissant autant que possible l'utilisateur s'exprimer comme il le souhaite.

Dans la section 2 nous présentons les stratégies mises en œuvre dans notre système de dialogue pour d'une part améliorer la navigation et d'autre part gérer au mieux les erreurs survenant en cours de dialogue. C'est en situation d'erreur que l'initiative doit être reprise par le système. La section 3 décrit quelques expériences et évaluations ayant permis de valider nos choix. Ces évaluations ont été faites, pour la plupart, dans le cadre des évaluations finales du projet ARISE menées par la SNCF (partenaire du projet) auprès du grand public en novembre 1998.

2. STRATÉGIES ET SYSTÈME DE DIALOGUE

Le terme **stratégies** recouvre différents aspects intervenant lors de la conception et de la mise en œuvre d'un système de dialogue oral homme-machine. L'aspect le plus théorique concerne la nature de l'interaction elle-même et reprend quelques-uns des aspects fondamentaux de la communication [Rou91, Aus62]. Il s'agit là de méta-stratégies qui vont avoir une influence sur des choix stratégiques ultérieurs. Le deuxième aspect dépend des choix ergonomiques fixés lors de la conception du système [Ros99]. Nous nous référons ici aux qualités dont nous voulons doter le système. Il s'agit donc de tout ce qui concerne la négociation, la navigation, la souplesse. Le troisième aspect, plus technique, concerne des choix de développement. Ces différents aspects sont fortement liés et dépendants des objectifs généraux fixés au départ, notamment en matière d'ergonomie et sont sous-tendus par les méta-stratégies. Nous rappelons donc nos objectifs pour ensuite décrire comment les atteindre en améliorant la structuration de la tâche et les capacités du gestionnaire de dialogue.

2.1. Les objectifs

Nous voulons d'abord un système de dialogue à initiative partagée permettant à l'utilisateur d'interagir aussi naturellement que possible avec la machine. Ce système doit permettre la navigation, la négociation et doit être capable de détecter et de gérer au mieux les erreurs. Ensuite nous voulons permettre aux utilisateurs confirmés d'aboutir avec un nombre minimal d'échanges.

2.2. Structuration du domaine

La structuration du domaine en tâches ou en fonctionnalités présuppose connaissance et organisation de ce do-

maine (les termes tâche ou fonctionnalité, a peu près synonyme ici, sont utilisés suivant qu'on s'intéresse plutôt au système ou plutôt à l'utilisateur).

Domaine désigne l'ensemble des "secteurs d'activités" possibles, **tâche** les différentes actions possibles côté système et leur organisation, **fonctionnalités** les différentes actions qu'on veut rendre accessibles à l'utilisateur. Pour un domaine donné définir la tâche revient à déterminer ce qu'un utilisateur doit pouvoir faire *avec* le système. Dans le cas d'ARISE le domaine est celui des trains voyageurs de la SNCF. L'utilisateur doit pouvoir accéder aux informations concernant les horaires des trains incluant les horaires de départ et d'arrivée et les changements le cas échéant. L'utilisateur doit pouvoir effectuer des réservations ou un achat de billets, s'informer sur les réductions ou les tarifs. Après avoir défini les différentes fonctionnalités, la structuration du domaine (le modèle de la tâche) peut être élaborée. Il s'agit de déterminer des relations (fortes ou faibles, hiérarchiques ou non) entre les différentes fonctionnalités, chacune pouvant être découpée en plusieurs buts à atteindre. Ces buts correspondent à des fonctionnalités élémentaires. Afin de rendre le dialogue aussi souple que possible c'est-à-dire de laisser un maximum de liberté à l'utilisateur notamment en matière de navigation et de négociation, nous avons distingué différents types de fonctionnalités (informations horaires, informations services, réservation, achat billet et informations réductions) et nous avons défini des fonctionnalités élémentaires lesquelles ne sont pas nécessairement affiliées à un type de fonctionnalité donnée mais plutôt à l'ensemble de la tâche. Le modèle de la tâche peut être vu comme un réseau dont les nœuds sont des fonctionnalités avec des liens entre ces nœuds si le passage entre fonctionnalités est possible.

2.3. *Le gestionnaire de dialogue*

Afin d'atteindre nos objectifs le gestionnaire de dialogue, dont le rôle est primordial, doit avoir connaissance de ce qui s'est passé et se passe au cours du dialogue, et surtout il doit être capable de prévoir afin de mieux conseiller ou, le cas échéant diriger l'utilisateur. Ces connaissances sont utilisées différemment selon l'état du dialogue pour lequel différentes phases sont distinguées.

Les connaissances Le gestionnaire dispose d'un ensemble de connaissances statiques (essentiellement connaissances linguistiques et modèle de la tâche) et dynamiques. Ces dernières, que nous décrivons, sont utilisées à différents moments de l'analyse de la requête. Elles participent à l'évaluation de l'état du dialogue. Ces diffèrentes sources de connaissances sont :

1- *Paire d'échange système-utilisateur:* Il s'agit du couple (parole système, parole utilisateur) courant, qui permet d'extraire certaines informations utiles pour la gestion à court terme du dialogue. Par exemple, si à une question du système proposant une réservation, l'utilisateur répond par l'affirmative, la fonctionnalité réservation sera activé.

2- *Historique du dialogue:* Il s'agit d'un historique sur un plus long terme que la simple paire système-utilisateur. Cet historique comprend trois niveaux:

- *Dialogue:* il s'agit de l'histoire du dialogue dans son ensemble, de la liste des tâches [1] ouvertes, et du bon ou mauvais déroulement du dialogue.
- *Tâche:* il s'agit pour une tâche donnée, de la suite des sous-tâches qui ont été appelées et de l'historique du bon ou mauvais déroulement de la tâche.
- *Sous-tâche:* il s'agit pour une sous-tâche donnée de son histoire c'est à dire de tout ce qui concerne son bon ou mauvais déroulement.

Cet historique va permettre de gérer la navigation, la négociation et le traitement des erreurs.

3- *État du dialogue:* Il est déterminé par la coexistence de certaines informations se rapportant tant à la tâche en cours qu'au déroulement du dialogue (durée, nombre de requêtes, présence d'une incohérence) et par les différentes sources de connaissances. L'état du dialogue permet une adaptation au dialogue en cours, un meilleur repérage des erreurs et une génération de réponse plus adaptée.

Les phases Une phase correspond à l'action qui est en cours dans une (sous-)tâche donnée (p. ex. la négociation dans la recherche d'un horaire). Cela se rapproche à un niveau plus théorique d'un acte de dialogue. Les différentes phases de dialogue identifiées sont l'acquisition, la négociation, la navigation, la post-acceptation et le méta-traitement.

Durant la phase d'**acquisition**, le système doit obtenir les informations nécessaires pour compléter la tâche ouverte. Celles-ci dépendent bien entendu du domaine d'application et du modèle de la tâche. Si l'utilisateur accepte la solution proposée par le système, alors le système entre dans une phase de **post-acceptation**. Dans cette phase, soit l'utilisateur fait une autre demande soit le système propose de lui-même d'autres demandes et ce en fonction du modèle de la tâche qui prend là toute son importance. Cette phase de post-acceptation est une phase de transition pour passer à une phase de **navigation** ou à une phase de **négociation** ou à la clôture du dialogue. La phase de **méta-traitement** concerne le repérage et le traitement des erreurs. Ces phases sont des états à partir desquels le système peut prendre des décisions.

Lors du dialogue, chaque énoncé de l'utilisateur passe par différentes étapes.

1- Compréhension contextuelle: il s'agit de l'interprétation du schéma sémantique issu du module de compréhension littérale dans le contexte du dialogue courant.

2- Décisions: il s'agit de décider s'il y a un changement de tâche (navigation), une erreur (p.ex. une contradiction entre l'historique du dialogue et l'interprétation contextuelle) ou si le dialogue suit son cours. Le changement de tâche implique une fermeture de la tâche en cours et une activation de la nouvelle tâche. Une trace du changement est conservée de façon à faciliter le retour à la précédente tâche. Si une erreur est détectée, la stratégie pour la traiter dépend de la phase de dialogue dans laquelle se trouve le système (acquisition ou post-acquisition), de la tâche courante, des historiques et du modèle de la tâche.

La navigation La navigation est directement dépendante du modèle de la tâche. Idéalement, nous pourrions envisager la navigation comme une liberté totale laissée à l'utilisateur de changer de fonctionnalité à n'importe quel moment du dialogue. Dans la réalité (pour ARISE en tout cas), le domaine est organisé de telle manière que seuls certains passages sont autorisés. Ceci nous permet d'utiliser au mieux la prédiction et limite les erreurs

[1] Les tâches et sous-tâches dont nous parlons maintenant sont l'équivalent système des fonctionnalités et fonctionnalités élémentaires décrites dans le cadre du modèle de la tâche.

d'interprétation même lorsqu'elles sont dues au système de reconnaissance. Le module de décision se fonde sur le modèle de la tâche pour décider de la validité de ces passages. Tant que le dialogue se passe à l'intérieur d'une des sous-tâches, telles que nous les avons décrites plus haut, la navigation, sauf cas exceptionnels, n'est pas autorisée. Par exemple, si le système en est encore à acquérir l'un des trois éléments nécessaires à l'accès à la base de données (ville de départ,ville d'arrivée et date de départ), une demande sur les réductions ou les tarifs ne pourra être satisfaite. En revanche le passage d'une sous-tâche *réservation* à *informations réduction* est libre.

Le traitement des erreurs Une erreur représente autant une contradiction due à l'utilisateur (dans une même requête ou entre différentes requêtes) qu'une erreur due au système lui-même. Pour le système, une erreur est surtout une incohérence ou une contradiction dans le temps. Nous distinguons deux catégories d'erreurs: les erreurs systématiques du système de reconnaissance et les autres. Les premières erreurs sont traitées et corrigées lors de la compréhension contextuelle avant même que l'énoncé ne soit analysé par le module de décisions. Ces erreurs sont fortement dépendantes des caractéristiques du système de reconnaissance. Le premier exemple de la figure 1 illustre bien ce type d'erreur: il s'agit ici de la confusion, fréquente, entre "non"et "Nantes". À ce moment-là du dialogue, si le système récupère "Nantes", il le transforme en "non". La nouvelle structuration de la tâche a rendu ce type de traitement possible et relativement performant. La détection d'autres types d'erreur repose sur un contexte plus large que la simple paire système-utilisateur. En cas de conflit détecté lors de la phase d'acquisition, le système passe d'un dialogue libre à un dialogue contraint. Lors du passage à un dialogue contraint le système de reconnaissance bascule vers un lexique et un modèle de langage plus spécifiques. Ce passage est visible pour l'utilisateur dans la mesure où le système lui demande de respécifier l'information contradictoire et lui fournit un exemple de formulation possible, comme le montre l'exemple 2 de la figure 1. Le dialogue contraint peut-être déclenché dans deux cas: le système rencontre une difficulté pour obtenir une des valeurs de base (ville de départ, d'arrivée et date) ou bien lorsque le système rencontre une valeur contradictoire, lors d'une autre phase, toujours à propos d'une des trois valeurs de base. Ce type de contradiction peut aussi bien être due à une erreur ancienne mais non détectée qu'à une erreur récente.

3. EXPÉRIENCES ET ÉVALUATIONS

Un corpus de plus de 3700 appels a été utilisé pour développer et tester les stratégies de dialogue. Le système a été régulièrement évalué par la SNCF (partenaire du projet ARISE). La dernière évaluation a eu lieu en novembre 1998. Les sujets ont été recrutés par une compagnie indépendante. Le taux de réussite était de 15.3% sur les 151 appels évalués.

Nous présentons ici deux résultats d'évaluation portant plus spécifiquement sur la navigation d'une part et d'autre part sur les stratégies de récupération et traitement des erreurs. Pour ces deux évaluations, quatre versions différentes du système ont été testées. Pour l'évaluation de la navigation (voir figure 2), nous avons utilisé deux versions (mai'98 et octobre'98). La première permettait de naviguer entre certaines fonctionnalités par simple détection de mots-clés dans la requête de l'utilisateur. La deuxième présente une plus grande connectivité entre fonctionnalités, mais les conditions sur les transitions sont plus fortes. Cette version inclut toutes les dernières modifications en matière de structuration de la tâche et de récupération des erreurs. Dans cette version, nous avons implémenté deux niveaux de dialogue (l'un libre et l'autre contraint). Le passage à un niveau contraint se fait lorsque le système détecte des erreurs ou lorsqu'il ne parvient pas à obtenir les informations dont il a besoin pour interroger la base de données. Pour évaluer le traitement des erreurs, nous avons utilisé les données recueillies lors de deux évaluations menées par la SNCF. Le premier système évalué (novembre 97) est la version de base de notre système. Le deuxième est le même que celui utilisé pour les évaluations d'octobre 98 mis à part le nombre de fonctionnalités implémentées. Dans cette version il y a ni réservation ni tarifs, en revanche il y a possibilité de demander des trajet aller-retour.

S: mercredi 19/04 vous avex un train direct de Paris à Lille à 9h10 arrivant à 10h08. Qu'en pensez-vous? Voulez-vous réserver?
R: je voudrais le tarif
S: Bénéficiez-vous d'une réduction ?
R: Nantes — U: non
S: En quelle classe voulez-vous voyager ?

R: je vais à Lyon
S: Oui, vous allez à Lyon de quelle ville partez-vous ?
R: je vais de Lyon à Lille
S: Je suis désolée, je n'arrive pas à comprendre, indiquez votre ville de départ, par exemple Paris.
R: Lyon
S: Bon, vous partez de Lyon quelle est votre destination ?

Figure 1: Exemples de détection et de traitement d'erreur. S: système; R: reconnaissance; U: utilisateur (si différence avec la reconnaissance)

3.1. *La navigation*

Nous avons comparé deux versions du système. La première version (mai'98), nous a permis de constater qu'une trop grande liberté de navigation (conditions trop faibles sur les transitions) limitait la capacité de prédiction pourtant nécessaire pour diminuer le risque d'erreurs. Durant l'automne 98, nous avons testé une version dans laquelle la tâche est structurée différemment ce qui permet d'utiliser au mieux la capacité de prédiction, en particulier en limitant les passages d'une tâche à une autre. La figure 2 donne un aperçu partiel des changements opérés dans la structuration du domaine.

Nous avons mesuré les réussites en terme de passage d'une tâche à une autre. Le gain global a été de 10% entre les deux versions (en mai 98: 82% de passages demandés ont réussi et 92% en octobre 98). Le tableau 1 montre le nombre de requêtes avec un changement de tâche et le nombre de fois où le changement portait sur le tarif, la réservation et les services (calculé sur 109 dialogues). Le gain est net en particulier sur la détection de la tâche pour les demandes de tarif.
Il nous fallait également vérifier si le système ne passait pas à une autre tâche alors que celle-ci n'était pas demandée. Entre mai 98 et octobre 98 le nombre de passages

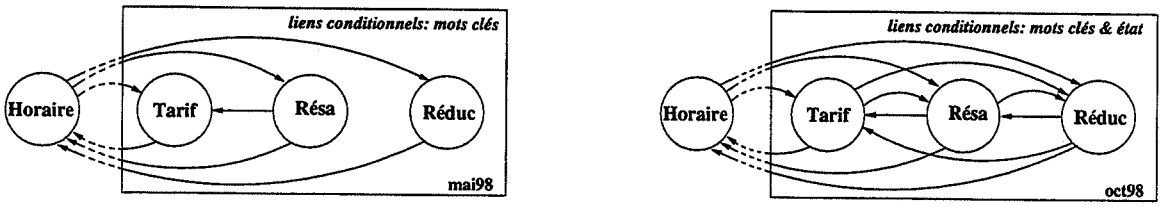


Figure 2: Extraits de la structuration du domaine. La partie gauche correspond au système mai'98, la partie droite au système octobre'98. Plus de possibilités de navigation sont offertes mais avec des conditions sur les arcs dépendant à la fois de la requête de l'utilisateur et de l'état du dialogue.

Table 1: Détection de changements de tâches: # changements demandés / # changements effectués.

Données	*Tarif*	*Résa*	*Serv.*
mai98 (50 dial.)	37/25	67/55	38/37
oct98 (59 dial.)	53/48	61/57	24/22

effectués alors que non demandés a diminué (de 40% des passages effectués qui n'étaient pas demandés en mai 98 à 21% en octobre 98).

3.2. *Traitement des erreurs*

Une façon simple de vérifier l'apport des stratégies de récupération d'erreurs (notamment celles dues à la reconnaissance, de loin les plus fréquentes) est de mettre en rapport le taux de succès de dialogue et le taux d'erreurs de reconnaissance. Nous avons comparé deux versions du système. La figure 3 nous montre un rapport entre le taux de succès du dialogue et le taux d'erreurs de reconnaissance. Nous constatons qu'à un taux d'erreurs de reconnaissance identique, le taux de réussite de dialogue a augmenté entre les deux versions du système (par exemple, pour 30% d'erreurs de reconnaissance, nous avions en 97 63% de succès dans le dialogue et 73% en 98). Nous pouvons également constater que plus le taux d'erreurs de reconnaissance est élevé plus le gain sur le taux de succès de dialogue est important. Ceci est dû à l'utilisation d'un dialogue à deux niveaux dont l'un (contraint) prend la relève de l'autre (libre) en cas de problèmes ainsi que la nouvelle structuration de la tâche (à la fois libre mais plus prédictive).

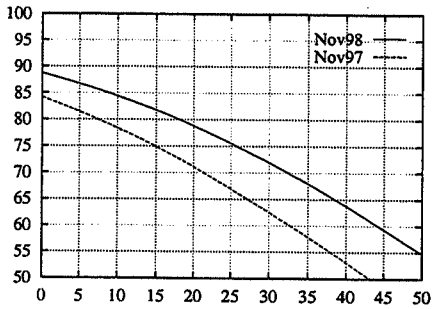


Figure 3: Taux de succès du dialogue en fonction du taux d'erreur de reconnaissance.

4. Discussions

Les stratégies regroupent en fait trois ensembles distincts qui interviennent à des moments différents lors du développement d'un système de dialogue. Ces trois ensembles recouvrent des contraintes méta-stratégiques, des contraintes ergonomiques et des contraintes techniques. Chaque décision prise à l'un de ces trois niveaux va influencer les choix possibles aux autres niveaux. Afin de parvenir à un dialogue peu contraint et permettant une navigation et une négociation aisée, nous avons amélioré la structuration du modèle de la tâche. Toutefois, il s'agit de garantir simultanément une grande liberté à l'utilisateur dans le dialogue et une bonne robustesse face aux erreurs. Laisser l'initiative de la conduite du dialogue majoritairement à l'utilisateur suppose alors une bonne capacité de prédiction sur les choix les plus probables de l'utilisateur. Pour cela, nous avons modifé non seulement le modèle de la tâche mais aussi le gestionnaire de dialogue.

Différentes expériences et évaluations nous ont permis de montrer qu'une navigation, une liberté contrôlées un modèle de la tâche prédictif amélioraient la qualité du dialogue tout en facilitant le traitement des erreurs.

Nous n'avons pas évalué de façon précise toutes les stratégies, notamment pour le traitement des erreurs. Pour le moment nous pouvons seulement dire que nous avons obtenu une amélioration globale. Quelle stratégie apporte le plus? Pour décider de cela d'autres évaluations sont nécessaires.

Bibliographie

[Rou91] E. Roulet, A. Auchlin, J. Moeschler, C. Rubattel, M. Schelling, "L'articulation du discours en français contemporain, 3ème édition, Sciences pour la Communication, Peter Lang,Berne, 1991.

[Aus62] J. L. Austin, "How to do thing with words", Oxford: Clarendon Press, 1962.

[Ehr99] U. Ehrlich, "Task hierarchies representing subdialogs in speech dialog systems", Eurospeech'99.

[Hua99] Chao Huang et al., "Lodestar: a mandarin spoken dialogue system for travel information retrieval", Eurospeech'99.

[Lam98] L. Lamel et al., "The LIMSI ARISE system", *IVTTA'98* pp. 209-214, Torino, Sept. 1998.

[Lav99] C. Alexia Lavelle et al., "Confirmation strategies to improve corrections rates in a telephonic inquiry dialogue system", Eurospeech'99.

[Os99] Els den Os et al., "Overview of the Arise project", Eurospeech'99.

[Ros99] S. Rosset, S. Bennacef, L. Lamel, "Design strategies for spoken dialog systems", Eurospeech'99.

Reconnaissance de la langue et du locuteur

Longueur de confusion sur la plage vocalique

Ben Kaehler, John Smith and Joe Wolfe

School of Physics
University of New South Wales, Sydney, Australie
Tél: (61) (2) 93854954 - Fax (61) (2) 93856060
J.Wolfe@unsw.edu.au - http://www.phys.unsw.edu.au/~jw/index.html

ABSTRACT

Subjects were asked to identify monosyllabic synthesized words. The formants of the synthesized vowels were known, so the choices of a subject produces a perceptual map of his/her language. We compare such maps with similar maps for the vowels produced during speech. The chance of identifying a sound as a particular vowel decreases with its normalized displacement from the mean position of that vowel in the vocal plane. A plot of probability of identification as a function of separation in this plane defines a characteristic resolution/confusion 'distance' for a particular language. We report results of an experiment to determine such characteristic distances using synthesized speech and an automated testing routine.

1. INTRODUCTION

La plage vocalique est continue mais sa division en phonèmes est quantique. La reconnaissance des voyelles est un des exemples les plus longtemps étudiés du processus de catégorisation perceptive sur un axe continu.

Le nombre de voyelles varie considérablement selon les langues. Un nombre plus grand donne, en principe, la capacité de transmettre l'information à une vitesse élevée, mais à condition que la taille de la discrétisation dans l'espace continu dépasse la sensibilité ou la résolution du récepteur. Nous avons proposé une façon de définir une résolution moyenne dans un cas simple [Dow97], et calculé des valeurs à partir d'un ensemble de phonèmes français. Les mesures de la résolution sont statistiques et donc fonction de la façon d'échantillonner l'espace vocalique. La synthèse des sons permet un échantillonnage de tous les paramètres, sous contrôle explicite de l'expérimentateur. Nous présentons ici des mesures de la catégorisation perceptive de voyelles synthétiques.

Nous ne considérons ici que des voyelles à l'intérieur de mots monosyllabiques. Ceux-ci sont des cas simples, parce que, dans la parole normale, la reconnaissance des voyelles dépend fortement du contexte lexical et grammatical. Maints exemples existent, néanmoins, où la résolution d'une seule voyelle dans un mot isolé est nécessaire pour la compréhension, en particulier quand il s'agit de noms propres.

1.1 Mots prononcés ou synthétiques ?

Dans une étude précédente [Dow97; Dow00], un jury de douze francophones ont écouté trois répétitions de 250 voyelles prononcées par les sujets dont les fréquences de résonance du conduit vocal étaient mesurées. Les membres du jury indiquèrent à quel mot français le son qu'ils venaient d'entendre ressemblait le plus.

Cette expérience avait les avantages suivants : (i) le son était naturel et (ii) les résonances était déterminées par un algorithme qui n'avait pas besoin du jugement humain (on n'avait pas besoin de repérage formantique). Son désavantage était le manque de contrôle sur l'échantillonnage de la plage vocalique : le groupement des sons autour des voyelles naturelles du français limitait l'expérience et influençait le choix vers ces régions. De plus, l'existence de régions vides sur la plage vocalique empêche l'investigation de la perception des sons dans ces régions par une expérience dont les stimulis sont des voyelles naturelles.

Pour cette raison nous faisons une deuxième expérience, où les membres de plusieurs jurys écoutent et classent une série de sons synthétisés avec une gamme de formants connus qui échantillonnent la plage en densité uniforme. Pour cette expérience, nous sommes en train d'étudier trois langues différentes (l'anglais, l'espagnol et le français) mais jusqu'à présent nous n'avons des résultats statistiquement significatifs que pour l'anglais.

Pour chaque langue, nous analysons le classement des voyelles afin de déterminer un déplacement caractéristique ou longueur de confusion sur la plage vocalique de perception dans les conditions de l'expérience.

1.2 Formants et résonances

Dans une première approximation, les voyelles des langues occidentales sont classées et caractérisées par les deux ou trois premiers formants (F1, F2, F3), même si la durée et le ton sont parfois importants. Nous réservons le mot 'formant' pour décrire un pic dans l'enveloppe spectrale du son d'une voyelle. Dans l'étude précédente, nous avons mesuré les résonances acoustiques (R1, R2) du conduit vocal qui produisait la voyelle. On s'attend à ce que les fréquences des formants soient approximativement égales à celles des résonances, mais des différences non négligeables peuvent être introduites par l'interaction source-conduit de la source glottale et par l'impédance de radiation [Fan73].

1.3 Plans de production et de perception

Dans les espaces (F1, F2, F3) ou (R1, R2, R3), on peut classer les points en voyelles selon l'intention du locuteur ou selon la perception de l'auditeur [Fan73, Lan77, Dow97]. Quand un sujet prononce une voyelle, on peut mesurer ses formants ou les résonances du conduit vocal pendant la production. Un ensemble de mesures constitue ce que nous appelons le plan de production des voyelles de cette langue (et de ce sujet). Pour créer ce que nous appelons le plan de perception ou plan perceptif, un sujet identifie, comme voyelle de sa langue, un son dont on connaît les formants ou un son produit par un conduit vocal dont on connaît les résonances.

D'habitude, la compréhension d'une langue parlée, et surtout des voyelles isolées, est le résultat d'une forte ressemblance entre le plan de production et celui de perception, mais la compréhension des accents divers et des cas spéciaux tels que la parole sous hélium montre que la ressemblance peut être relative plutôt qu'absolue. Nous rapportons ici quelques différences quantitatives et qualitatives entre ces deux plans.

1.4 Distances sur la plage

Que veut dire déplacement sur la plage vocalique ? La plage de variation de R2 ou F2 est plus grande que celle de R1 ou F1. Pour cette raison, nous avons défini [Dow97] un déplacement sans dimension entre les points *a* et *b* sur la plage en deux dimensions comme suit:

$$d \equiv \sqrt{\left(\frac{R1_a - R1_b}{\sigma_1}\right)^2 + \left(\frac{R2_a - R2_b}{\sigma_2}\right)^2} \qquad (1),$$

où σ_1 est l'écart type mesuré pour tous les résonances R1 de la langue, σ_2 est celui des R2. Dans l'expérience rapportée ici, les formants remplacent les résonances dans l'équation (1).

1.5 Résolution et longueur de confusion

Si on déplace une voyelle de sa place moyenne sur la plage vocalique, *quel sera le déplacement caractéristique moyen à partir duquel les auditeurs commencent à la confondre avec une autre ?* Pour chaque voyelle, on peut trouver ($\overline{F1}$, $\overline{F2}$) ou ($\overline{R1}$, $\overline{R2}$), le point du plan (F1,F2) ou (R1,R2) qui a les valeurs moyennes de tous les sons reconnus comme cette voyelle. Ce point est appelé la "place moyenne". Plus on est loin de ce point, plus grande est la chance de reconnaître le son comme une autre voyelle, ou comme un son qui n'est pas une voyelle. Nous avons trouvé [Dow97,Dow00] que la probabilité de reconnaissance d'un son (R1,R2) comme une voyelle ($\overline{R1}$, $\overline{R2}$) diminue de façon exponentielle avec la distance entre ces points. Le calcul de la longueur de confusion est compliqué parce que les mesures statistiques d'identification sont fonction de l'échantillonnage de la plage vocalique.

2. METHODOLOGIE

2.5 Mots synthétiques

Les voyelles sont placées dans des mots monosyllabiques, choisis pour chaque langue afin de minimiser le nombre de mots sans sens. En anglais, la forme 'h<V>d' ne donne qu'un seul mot sans sens en anglais[1]. Pour les voyelles <V>, F1 et F2 furent altérés par pas de 5% sur un quadrilatère ((300,700) (700,1000) (1450,700) (1950,300) (valeurs en Hz)) du plan (F2,F1) et F3 a pris quatre valeurs entre 2.1 et 3.1 kHz. L'enveloppe spectrale a été calculée explicitement et les voyelles ont été synthétisées par somme de sinus. Les fréquences fondamentales moyennes sont respectivement 132 et 250 Hz pour les voix "masculines" et "féminines", avec une réduction légère et cubique pendant le mot. Les valeurs (en %) du *flutter, creak, jitter* et *shimmer* sont (0.5, 0.5, 1.5, 1.5) et (0, 0.2, 0, 0) pour les deux. La largeur de bande des formants a été déterminée par la régression polynomiale décrite par [Haw95], et la taille par régression linéaire sur les fréquences centrales des formants. Le /h/ est du bruit blanc, caractérisé dans le domaine spectral pour des échantillons tous les 10 Hz et passé par les formants de la voyelle suivante. Les locus pour le /d/ sont 1749 Hz et 2000 Hz et un délai de 50 ms précède une plosive de 30 ms.

2.6 Classement des sons synthétiques

Le système de classement est automatisé. Les sujets répondent aux questions posées par un ordinateur qui synthétise les 'mots' en ordre aléatoire pour chaque auditeur. Chaque mot est produit trois fois à la suite, et un sous-ensemble d'approximativement 100-130 mots décrivant l'espace paramétrique (F0,F1,F2,F3) est présenté dans un ordre aléatoire pour chaque sujet. Pour l'expérience sur l'anglais, 113 volontaires, étudiants de l'Université de Nouvelle Galles du Sud à Sydney, ont participé. Ils cliquent sur des 'boutons' nommés *had, hard, head, herd, heed, hid, hoard, hod, hood, hud, who'd* et *none of these*. Une autre série de boutons leur permet de classer le son sur cinq niveaux entre « *quite unnatural* » et « *quite natural* ». Pour les autres langues, nous sommes toujours en train de faire des mesures.

3. RESULTATS ET DISCUSSION

Les auditeurs anglophones ont donné aux mots synthétisés avec une voix masculine (c'est à dire de faible F0) une valeur moyenne de 4.8 sur la gamme *quite unnatural* (1) à *quite natural* (5). Les commentaires de certains membres du jury après les séances de mesure indiquent qu'ils ne

[1] 'heed' [i], 'hid' [ɪ], 'head' [ɛ], 'had' [æ], 'hard' [ɑ], 'hod' [ɒ], 'hoard' [ɔ], 'hood' [ʊ], 'who'd' [u], 'hud' [ʌ] et 'herd' [ɜ]. Malgré l'acronyme HUD (Head-Up Display), 'hud' n'est pas encore un mot anglais.

s'étaient pas rendus compte que les mots avaient été synthétisés.

3.1 Plan de perception

On a calculé pour chaque voyelle $(\overline{F2}, \overline{F1})$, les moyennes des valeurs de chaque son identifié comme cette voyelle. Les positions sur le plan perceptif des voyelles anglaises (Fig. 1) ressemblent globalement aux positions sur les plans de production en (F1,F2) [Ber67] et en (R1,R2) [Epp97]. La similarité entre /i/ et /ɪ/ n'est pas surprenante: en production, ces voyelles sont largement distinguées par longueur (/ɪ/ est courte). Le plan de perception des voyelles "masculines" est déplacé vers l'origine par rapport au plan de voyelles "féminines" (c'est à dire de F0 élevée non présentées). Ce déplacement rappelle le déplacement bien connu pour le plan de production.

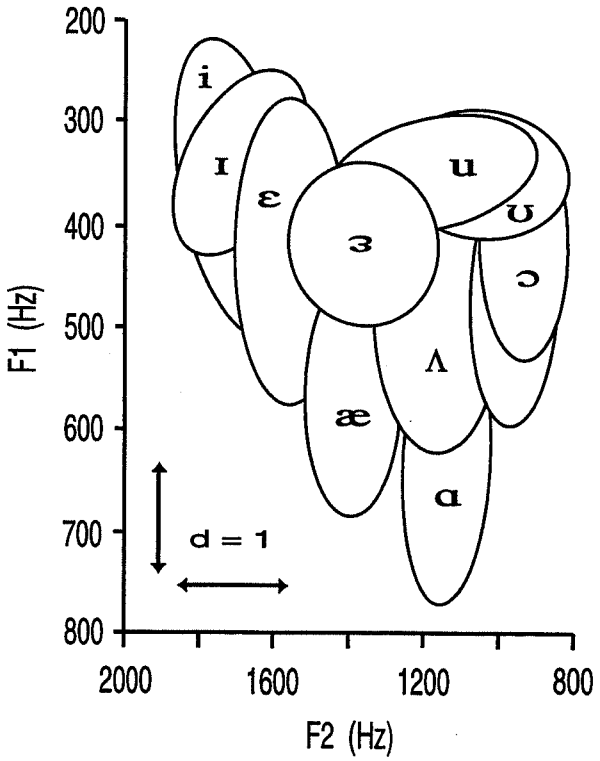


Figure 1 Plan perceptif des voyelles non nasalisées 'masculines', synthétisées pour échantillonner le quadrilatère vocalique en densité uniforme. Dans cette expérience, les auditeurs sont anglophones. Les axes sont proportionnés de telle sorte que les barres, qui indiquent les écarts types (σ_1 et σ_2) pour tous les R1 et R2, soient de longueurs égales. Ces barres sont donc de longueur d = 1, selon la définition (1). La pente du grand axe est le coefficient de régression linéaire et les axes des ellipses indiquent l'écart type pour chaque voyelle dans cette direction et la direction perpendiculaire.

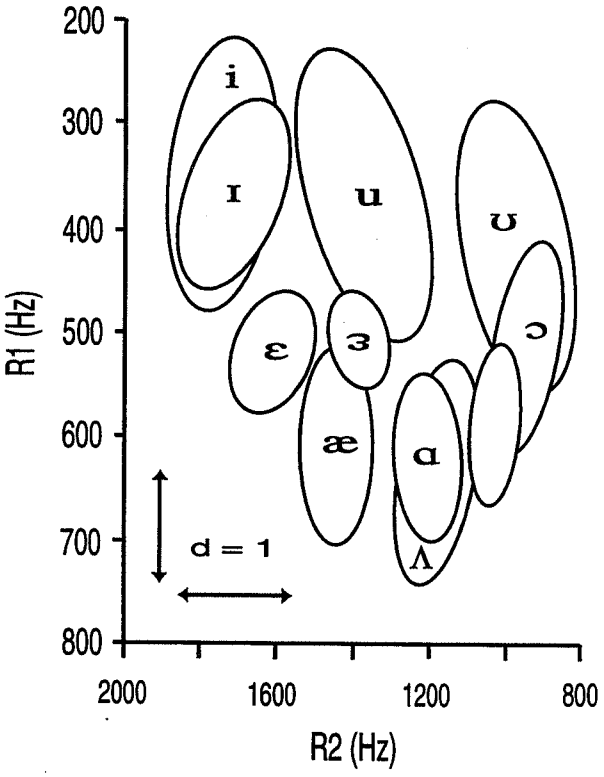


Figure 2 Le plan de production pour les voyelles de l'anglais australien, calculé à partir des données de [Epp97]. Mesures des résonances (R1,R2) des conduits vocaux de 33 jeunes australiens masculins lorsqu'ils prononcent les voyelles indiquées.

Dans l'expérience que nous rapportons ici, le quadrilatère vocalique est échantillonné partout, et le recouvrement de la plage de perception montrent que, même si l'anglais ne se sert pas de certaines régions de la plage, les anglophones reconnaissent des sons dans ces régions comme des voyelles. 12% des choix était '*none of these*', mais la distribution de tels choix couvre le plan entier.

3.2 Comparaison des plans de production et de perception

Pour ces sons synthétiques, la distribution de points pour une voyelle perçue est typiquement plus grande que sa distribution pour la même voyelle produite (voir Figures 1 et 2.). (Cette observation est aussi vraie pour le français, et nous avons mis ces résultats sur le site internet [Dow00].) La forme des distributions est différente aussi: dans le plan de perception les distributions sont plus larges (c'est à dire que le rapport $\sigma(F2)/\sigma(F1)$ pour une voyelle est typiquement plus grand que le même rapport dans le plan de production). Les positions relatives sur les deux plans sont approximativement semblables mais quelques différences sont évidentes. Sur le plan de production, [ʌ] (*hud*) et [ɑ] (*hard*) ont presque la même moyenne: ces deux voyelles sont distinguées en production largement par la durée ([ʌ] est courte). Pour

[ʌ] perçue, F1 est inférieure à sa valeur dans le plan de production, pour [ɑ] perçue, F1 est supérieure à sa valeur dans le plan de production.. Sur le plan de production, il y a une région (entre [u] et [ʊ]) dont les locuteurs anglophones ne se servent que rarement (Figure 1). Si F1 est faible, les sons dans cette région vide sont souvent perçus comme [u] ou parfois [ʊ]. Pour F1 plus important, ils sont perçus comme [ʌ].

Cette région du plan de production (F2 ~ 1200 Hz, F1 < 700 Hz) est vide en français aussi où elle fait partie du triangle nasal [Lan77;Dow00]. Dans notre étude précédente, pourtant, l'emploi de voyelles prononcées comme stimulus nous a empêché de mesurer cette région du plan de perception [Dow97].

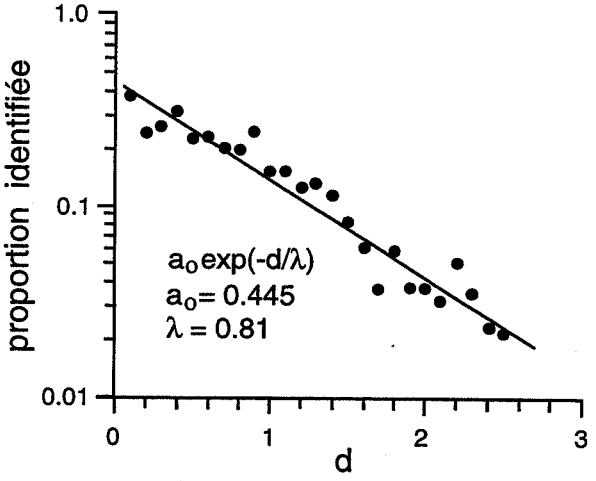


Figure 3 La proportion d'identifications en fonction de *d* sur le quadrilatère vocalique échantillonné en densité uniforme par des sons synthétiques.

3.3 Longueur de confusion

Pour chaque voyelle de place moyenne $(\overline{F2}, \overline{F1})$ le déplacement de chaque son de formants (F2,F1) a été calculé, et le nombre d'identifications a été compté pour chaque anneau $j.\delta \leq d \leq (j+1).\sigma$, $\delta = 0.05$. La Figure 3 montre la proportion identifiée dans chaque anneau en fonction de son rayon moyen. La figure montre aussi une fonction déduite des données: une exponentielle simple de longueur λ.

Pour l'anglais sous ces conditions $\lambda = 0.81$, c'est à dire que sa longueur de confusion est du même ordre mais est légèrement plus grande que l'écart type de tous les formants (en unités sans dimension, $\sigma(R1) \equiv 1 \equiv \sigma(R2)$). La Figure 1 et cette observation montrent que le plan de perception n'est pas divisé avec une grande précision. Le degré de recouvrement entre régions de voyelles contiguës est important. Hors de tout contexte, donc, on attend une confusion des voyelles voisines. La longueur de confusion est aussi du même ordre que F0. (λ est sans dimension. Sa taille en Hz est fonction de l'orientation. Dans la direction R1, $\lambda = 0.6$ F0. Dans la direction R2, $\lambda = 1.7$ F0.) Ce résultat n'est pas surprenant non plus : dans le domaine fréquentiel, l'enveloppe spectrale est échantillonnée par pas de F0, et on s'attend à ce que le limite de la résolution perceptive des formants soit de cet ordre, en l'absence d'autres informations.

3.4 Étude en cours

Il serait intéressant de pouvoir comparer des mesures des longueurs de confusion de langues ayant des nombres différents de voyelles. Pour cette raison, nous faisons actuellement des mesures automatisées en (F0,F1,F2,F3) pour le français (16 voyelles) et l'espagnol (5 voyelles), sous un protocole standardisé sur la forme décrite ici pour l'anglais. Cette étude nous donnera aussi (pour le français) la comparaison des méthodes des stimulus synthétiques et naturels.

Remerciements.

Nous remercions l'Australian Research Council qui a subventionné, en partie, cette recherche. Nous remercions aussi tous nos volontaires.

BIBLIOGRAPHIE

[Ber67] Bernard, J.R.L. (1967), "Length and identification of Australian vowels", Australasian Universities Modern Language Assoc. Vol 27, pp. 100-120.

[Dow97] Dowd, A., Smith, J.R. and Wolfe, J. (1997). "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time." Language and Speech, Vol 41, pp. 1-20.

[Dow00] Dowd, A., Smith, J.R. and Wolfe, J. (1900). "French vowels" http://www.phys.unsw.edu.au/~jw/french.html

[Epp97] Epps, J., Dowd, A., Smith, J.R. and Wolfe, J. (1997) "Real time measurements of the vocal tract resonances during speech", Eurospeech'97, pp. 721-724. www.phys.unsw.edu.au/~jw/Eurospeech.html

[Epp97] Epps, J., Smith, J., et Wolfe, J. (1997) "A novel instrument to measure acoustic resonances of the vocal tract during speech", Measurement Science and Technology, Vol. 8, pp. 1112-1121.

[Fan73] Fant, G. (1973), "Speech Sounds and Features", MIT, Cambridge, Mass.

[Haw95] Hawks, J.W. et Miller, J.D. (1995). "A formant bandwidth estimation procedure for vowel synthesis." J. Acoust. Soc. Am., 97, pp. 1343-1344.

[Lan77] Landercy, A. et Renard, R (1977). "Éléments de Phonétique." Didier, Bruxelles.

Détermination expérimentale d'indices linguistiques pour la discrimination des langues romanes

Ioana VASILESCU, Jean-Marie HOMBERT, François PELLEGRINO

Dynamique du Langage - I.S.H.

14, avenue Berthelot
69363 Lyon Cedex 07 – France

ABSTRACT

This paper deals with perceptual identification and differentiation of five Romance languages, namely French, Italian, Spanish, Portuguese and Romanian. Previous studies have investigated human capability to identify spoken samples in unknown languages after a relatively brief exposure. Accordingly, we conduct an analysis to determine which perceptual categories are salient in Latin languages identification. Three groups of listeners with different mother languages (French, Romanian and Japanese) have been considered. Results reveal that: - identification scores are a function of previous exposure to the languages, - the patterns used for the discrimination within the Latin family are mother tongue dependent and - the segmental cues emerging from the subjects' responses may be relevant in automatic language identification.

1. INTRODUCTION

L'identification automatique des langues a dernièrement bénéficié d'une approche complémentaire issue de l'étude de la discrimination linguistique par des êtres humains. Les résultats témoignent d'une capacité remarquable des auditeurs naïfs à identifier et discriminer des langues complètement inconnues après une période d'apprentissage réduite. Ainsi, les expériences de [Mut94a] ont mis en évidence le fait que même sans exposition préalable à une langue étrangère, les sujets humains sont capables de discrimination. Des taux de 80 % de réussite après une courte écoute (2s) sont rapportés dans [Sto97].

L'information portée par le signal sonore est complexe, et la perception humaine semble capable d'en extraire des traits spécifiques d'une langue à l'autre, en termes d'inventaire phonémique, de règles phonotactiques, de structure intonative ou syllabique. L'étude de la perception des langues par les êtres humains permet non seulement l'amélioration de la compréhension des processus linguistiques de phonologisation, mais fournit également de nouveaux traits pertinents dans le cadre des systèmes automatiques d'identification des langues. Les systèmes de reconnaissance actuels sont essentiellement basés sur la modélisation phonotactique et il semble évident que de nouveaux traits soient nécessaires afin d'améliorer les taux de réussite.

Plusieurs facteurs ont été analysés au travers des expériences en identification des langues par les sujets humains. Ils concernent le matériel linguistique, la structuration de la tâche ou les caractéristiques des sujets en termes de connaissances linguistiques préalables. Le facteur "matériel linguistique" a permis la mise en évidence de stratégies perceptuelles face un nombre important (10) de langues de test [Mut94b]. La variation des tâches (discrimination vs. évaluation de la proximité des langues) a permis l'émergence de traits linguistiques discriminants [Sto96]. Le facteur "population" a révélé des différences dans la discrimination des langues en fonction du temps d'apprentissage ou du caractère monolingue ou bilingue des sujets [Mar99].

Notre étude se propose de vérifier dans quelle mesure la langue maternelle des sujets participants au test influence le type d'information linguistique que ces derniers vont utiliser pour la discrimination de 5 langues romanes, à savoir le français, l'italien, l'espagnol, le roumain et le portugais. Trois populations ont participé à l'expérience : deux d'entre elles étaient constituées de locuteurs natifs d'une langue romane (Français et Roumains) et la troisième, ayant le rôle de population de contrôle était constituée de sujets Japonais ne possédant aucune exposition préalable aux langues romanes.

Nous allons par la suite procéder à une brève présentation de la famille des langues romanes et des traits segmentaux et supra-segmentaux *a priori* pertinents pour leur discrimination. Dans le Section 3 nous présentons le corpus linguistique, la tâche de discrimination et les caractéristiques des populations participantes. Finalement, la 4ème Section décrit les résultats et fournit leur interprétation.

2. LES LANGUES ROMANES

Les langues romanes sont les idiomes issus du latin, un dialecte italique faisant partie de la branche occidentale de la famille indo-européenne, à savoir la branche italo-celtique [Ruh91]. La famille romane réunit 5 des langues les plus utilisées dans le monde contemporain. Elles représentent la langue ou l'une des langues officielles dans 7 pays européens (Belgique, France, Espagne, Portugal, Suisse, Italie et Roumanie), un continent (Amérique du Sud) et d'autres régions du monde (Canada, Amérique centrale etc.).

Notre approche prend en considération 5 langues romanes : le français, l'italien, l'espagnol, le portugais et le roumain. Ces langues sont représentatives de la famille romane du point de vue économico-politique, mais aussi de la distribution géographique (langues romanes occidentales : le français, l'italien, l'espagnol et le portugais vs. orientales : le roumain) qui entraîne des particularités spécifiques dues à des interactions avec d'autres langues (germaniques, slaves etc.). Une description phonologique de ces langues révèle en même temps que des traits communs génétiquement explicables, de nombreuses particularités qui concernent à la fois le niveau segmental et supra segmental.

La structure des systèmes vocaliques partage les langues romanes en deux groupes : celui des langues possédant deux oppositions vocaliques, antérieur/postérieur (italien, espagnol) et celui des langues possédant trois oppositions vocaliques, antérieur/postérieur/central ou antérieur-arrondi (roumain, français, portugais). De plus, la famille romane inclue deux des quatre langues européennes possédant des voyelles nasales phonologiques, moyennes et ouvertes (les deux autres langues étant le polonais et certains dialectes bretons) [Ruh74]. Les systèmes consonantiques sont plus homogènes en termes de traits communs, néanmoins des segments spécifiques témoignent des évolutions individuelles (par exemple, la consonne fricative glottale /h/ en roumain, la fricative dentale /θ/ en espagnol, etc.).

L'analyse du niveau supra segmental partage les langues romanes en 4 groupes en fonction de leur structure rythmique: syllabiques (l'italien et le roumain), accentuelle (le portugais), "trailer-timed" (l'espagnol) et à accent fixe ou "langue de frontière" (le français) [Hir98].

3. MATERIEL ET METHODE

3.1. Corpus linguistique

Les enregistrements (22kHz, chambre insonorisée, intensité normalisée) de 4 locuteurs, 2 hommes et 2 femmes, pour chacune des langues ont été utilisés pour cette expérience. Deux d'entre eux (homme et femme) ont été utilisés dans la phase d'apprentissage, les deux autres ayant servi à la constitution des stimuli de test. Le corpus inclut de la parole lue et des histoires quasi spontanées.

3.2. Populations

20 Français, 20 Roumains et 20 Japonais, hommes et femmes, âgés de 18 à 60 ans et ayant au moins un niveau de formation correspondant au baccalauréat, ont participé au test. L'exposition préalable aux langues de test est homogène pour chacun des groupes :

Les Français ont étudié l'espagnol à l'école, mais aucun d'entre eux ne parlait couramment cette langue ou aucune des autres langues romanes. De plus, la France est géographiquement en contact avec l'Espagne et l'Italie.

Les Roumains ont étudié le français à l'école, mais aucun d'entre eux ne le parlait couramment ou aucune autre langue romane. La Roumanie n'est voisine d'aucun des pays de langues romanes. Cependant des fictions télévisées produites en Amérique latine sont souvent diffusées sur les chaînes nationales en version originale sous-titrée.

Les Japonais n'ont étudié aucune des langues romanes et aucune exposition préalable à ces langues n'a été mentionnée.

3.3. Conditions d'expérimentation

L'expérience a été divisée en trois phases :

- L'entraînement a permis aux sujets de se familiariser avec chacune des langues romanes. Il a consisté dans l'écoute de deux extraits de 10s dans chaque langue. Les extraits, prononcés par deux locuteurs différents, un homme et une femme, ont été présentés en ordre aléatoire.
- Durant le test proprement dit, les sujets devaient prendre une décision de type "même langue"/ "langue différente" pour chaque item. 50 stimuli de type AB ont été présentés : Chaque stimulus durait en moyenne 6s et était séparé du second par un court son de type « cloche ». Le sujets disposaient de 2s après chaque séquence AB pour répondre si A et B provenaient de la même langue ou de langues différentes. Les extraits étaient présentés une seule fois et chaque combinaison L_i-L_j, où $\{i,j\} \in [1,\ldots,5]^2$ a été présentée deux fois.
- A la fin du test les sujets ont eu la possibilité de s'exprimer sur la nature des indices qui les ont aidé à discriminer les langues.

4. RESULTATS

4.1. Analyse préliminaire de la significativité

Un test de significativité statistique des réponses des trois populations a été effectué avant de procéder à leur analyse multidimensionelle. L'histogramme présenté dans la Figure 1 reproduit le pourcentage de réponses correctes pour chaque paire de langues, ainsi que les paires pour lesquelles les réponses ont été significatives ou non (t-test univarié, $p<0,001$). Les réponses des populations française et roumaine ont été significatives dans la plupart des cas (sauf les paires Portugais/Portugais pour les deux populations et Roumain/Portugais et Portugais/Roumain pour la population française). Les réponses des Japonais ont été données dans la majorité des cas au hasard. Par conséquent une analyse multidimensionelle de leurs réponses n'est pas pertinente ; elle n'est donc pas présentée dans la suite.

4.2. Analyse multidimensionnelle

Les analyses multidimensionnelles ont été effectuées à l'aide du logiciel Vista [Vis99] pour les réponses des populations française et roumaine.

Les sujets français. Les réponses ont reçu une représentation tridimensionelle. La Figure 2 montre la projection des résultats dans deux plans définis, le premier (D1/D2) par les deux dimensions principales, et le seconde par la première et la troisième dimension (D1/D3). Dans le plan D1/D2 trois groupes de langues apparaissent : la langue maternelle, les langues familières et les langues inconnues. La première dimension sépare la langue maternelle (français) des autres idiomes, tandis que la seconde permet de distinguer entre langues familières (italien, espagnol) et inconnues (portugais, roumain). La troisième dimension distingue plus nettement entre les langues familières, tandis qu'entre les langues inconnues la confusion se maintient. Il semble, par conséquent, que les sujets français soient difficilement capables de distinguer entre deux langues inconnues après une période très courte d'apprentissage. Finalement, des considérations phonologiques devraient aussi être prises en considération, dans la mesure où la seconde dimension semble séparer les langues en fonction de la complexité de leurs systèmes vocaliques : les langues à trois oppositions vocaliques (portugais, roumain) sont séparées des langues à deux oppositions (espagnol, italien).

Les sujets roumains. L'analyse des réponses à reçu également une représentation tridimensionnelle (Figure 3). La première dimension distingue la langue maternelle des autres langues romanes. La seconde dimension sépare les langues articulées autour de deux oppositions (italien, espagnol) des langues dont le système s'organise autour de trois oppositions (français, portugais). Le plan D1/D3 semble correspondre à une distribution géographique des langues, isolant les langues ibériques (espagnol, portugais) des autres langues romanes. Elle pourrait être également la conséquence de l'exposition fréquente à l'espagnol et portugais sud-américains au travers les fictions télévisées. La troisième dimension séparerait les langues familières (espagnol et au portugais) des langues moins familières (français et surtout italien qui n'est pas appris à l'école).

Les sujets japonais. Ils ont répondu au hasard, confirmant ainsi qu'une exposition préalable aux langues facilite grandement la tâche de discrimination. Des expériences impliquant un apprentissage plus long sont envisagées .

5. PERSPECTIVES

La présente étude visait à étudier les traits discriminants entre 5 langues romanes au travers d'une expérience perceptuelle. Elle met en évidence des stratégies différentes de discrimination des langues inconnues. La perception des langues étrangères semble être filtrée par les traits de la langue maternelle et repose sur différents types d'informations, linguistique (segmentale et supra segmentale) et/ou extra linguistique (notamment socio-linguistique pour ce qui est des sujets roumains qui doivent leur connaissances sur les langues ibériques aux médias). De plus, une exposition antérieure aux langues romanes et la connaissance d'au moins une de ces langues sont des facteurs fondamentaux dans leur discrimination. Les sujets réellement naïfs (les Japonais) sont incapables d'extraire des paramètres saillants après une courte exposition à la langue (20s).

Nos prochaines démarches tenterons de mieux définir les niveaux linguistiques responsables de la discrimination, de donner une meilleure définition à la notion d'"exposition à la langue" et, finalement, de valider les indices mis en évidence par l'approche expérimentale dans un système de reconnaissance automatique.

6. REMERCIEMENTS

Les auteurs remercient Sumikazu Nishio pour son aide dans la mise en place du protocole expérimental auprès des sujets japonais.

BIBLIOGRAPHIE

[Hir98] Hirst D., DiCristo A. (1998), Intonation Syste. A Survey of Twenty Languages, Cambridge University Press.

[Mar99] Marks E.A., Bond Z.S., Stockmal V. (1999) "The effect of proficiency in a specific foreign language on the ability to identifiy a novel foreign language", 14th ICPhS, pp. 133-135.

[Mut94a] Muthusamy I.K., Barnard E., Cole R.A. (1994), " Automatic language identification: A review/Tutorial", IEEE Signal Processing magazine.

[Mut94b] Muthusamy I.K., Jain N., Cole R.A. (1994), " Perceptual benchmarks for automatic language identification ", IEEE ICASSP.

[Ruh74] Ruhlen M. (1995), " Some comments on vowel nasalization in French. Notes and discussion ", Journal of Linguistics,10.

[Ruh91] Ruhlen M. (1995), Guide to the World's Languages, Stanford University Press.

[Sto94] Stockmal V., Muljani D., Bond Z. (1994), " Can children identify samples of foreign languages as same or different?", Language Sciences, 16, pp. 237-251.

[Sto96] Stockmal V., Muljani D., Bond Z. (1996), " Perceptual Features of Unknown Foreign Languages as Revealed by Multidimensional Scaling", ICSLP.

[Vis99] The Visual Statistics System Vista web page http://forrest.psych.unc.edu/research/ (visitée en Novembre 1999)

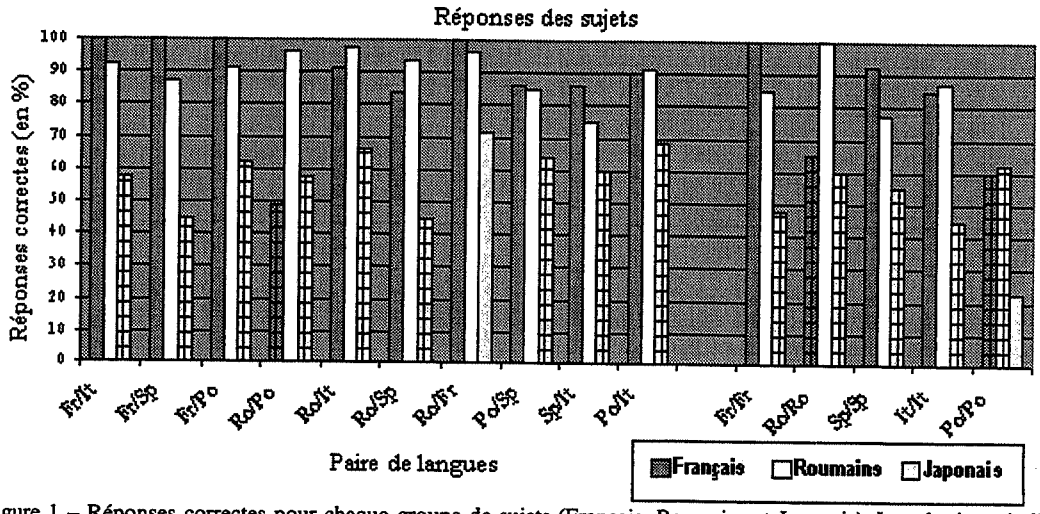


Figure 1 – Réponses correctes pour chaque groupe de sujets (Français, Roumains et Japonais). Les abscisses indiquent les paires de langue AB (AB et BA sont cumulés). Une barre pleine (resp. rayée) indique un score significatif (resp. non significatif) avec $p < 0{,}001$.

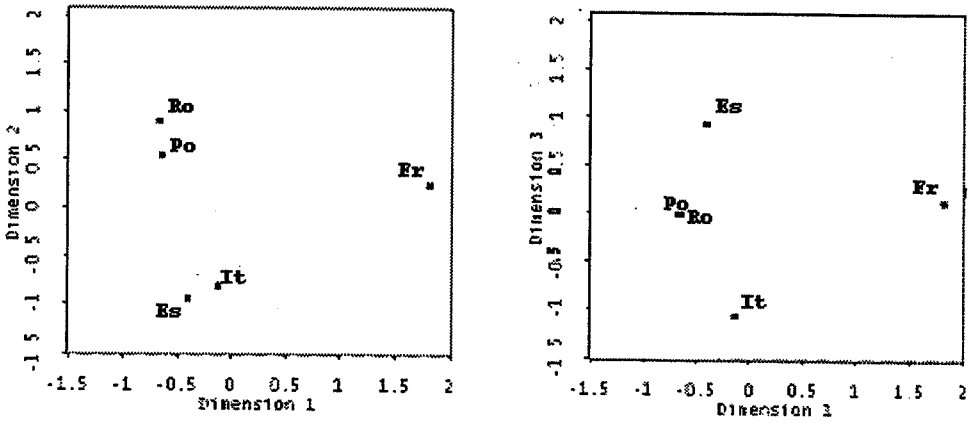


Figure 2 – Projection des réponses des sujets **Français** dans un espace multidimensionnel. A gauche la projection est réalisée dans le plan des deux premières dimensions et à droite dans le plan Dimension 1 / Dimension 3.

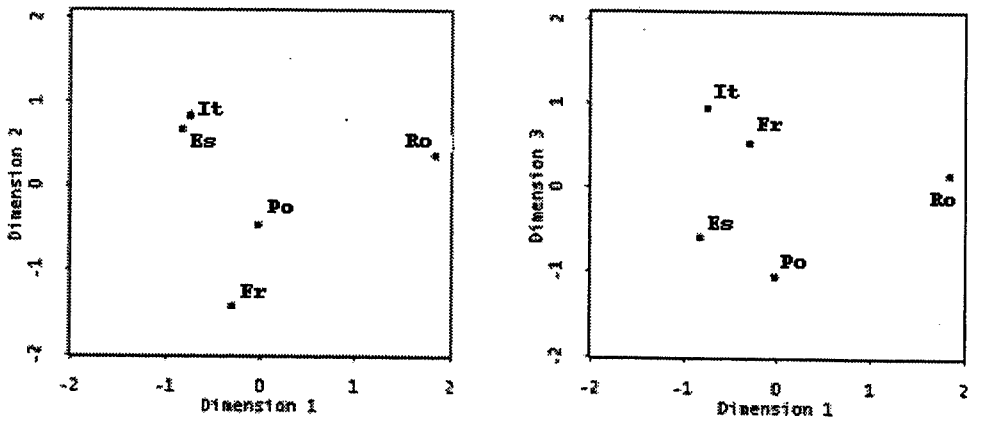


Figure 3 – – Projection des réponses des sujets **Roumains** dans un espace multidimensionnel. A gauche la projection est réalisée dans le plan des deux premières dimensions et à droite dans le plan Dimension 1 / Dimension 3.

Améliorer les systèmes de vérification automatique du locuteur en intégrant la variabilité émotionnelle : Méthodes et premières données

T. Bänziger, G. Klasmeyer, T. Johnstone, T. Kamceva, K. R. Scherer[1]

FPSE, Université de Genève
40 bv du Pont d'Arve, 1205 Genève
Tél. : +41 22 705 92 07 - Fax : +41 22 705 92 19
e-mail : banziger@fapse.unige.ch

ABSTRACT

Based on theoretical and empirical considerations it is claimed that sustained improvements in the accuracy of ASV algorithms can only be obtained by training on the basis of corpi that contain attitudinal and emotional speech variations. The results of two pilot studies with Swedish and German speakers provide preliminary support for this claim. A major ongoing study with French, German, and English speakers is described.

1. INTRODUCTION

Durant les dernières années un intérêt grandissant pour les systèmes de vérification automatique du locuteur (automatic speaker verification - ASV) s'est développé. De nombreux efforts ont été investis dans l'amélioration des modèles mathématiques et statistiques utilisés dans ce domaine. La plupart des systèmes actuels utilisent toutefois des paramètres directement empruntés à la technologie de reconnaissance de la parole qui, bien que similaire par certains côtés à la technologie de vérification du locuteur, se base à l'origine sur la recherche de paramètres qui minimisent la variabilité interlocuteur. Ces systèmes restent donc encore extrêmement sensibles aux changements transitoires de la voix du locuteur liés à des modifications provisoires de son état interne (fatigue, émotions, santé). En particulier, les modifications de propriétés vocales associées aux légères mais fréquentes variations de l'état émotionnel du locuteur (irritation, stress, satisfaction) sont susceptibles d'entraîner une augmentation du nombre de rejets de locuteurs qui devraient en principe être reconnus par le système (faux-rejets).

La prise en considération de ces variations devraient permettre d'améliorer la performance des systèmes ASV en diminuant le nombre de faux-rejets sans augmenter la quantité de fausses-acceptations (reconnaissance d'imposteurs par le système). Nous avons adopté deux angles d'approche complémentaires visant à atteindre cet objectif :

1. Améliorer les techniques d'enregistrement et la qualité et variabilité des enregistrements utilisés pour l'entraînement de ces systèmes.
2. Définir un ensemble de paramètres acoustiques plus optimaux.

Les premières étapes et résultats d'une recherche initiée dans cette optique seront décrits dans ce qui suit.

2. MÉTHODES

2.1 Modifier les modèles de locuteurs des systèmes ASV

Dans le cadre d'une approche visant à améliorer la performance d'un système ASV en modifiant l'entraînement du système, la première difficulté réside dans l'enregistrement d'une base de données incluant systématiquement des modifications au niveau de la voix/parole liées à des variations de l'état émotionnel et du degré de stress pour un grand nombre de locuteurs. L'un des objectifs lors de la construction d'une telle base de données est d'inclure un ensemble de variations vocales liées à des modifications subtiles de l'état du locuteur qui sont susceptibles de survenir fréquemment lors des interactions avec les systèmes ASV. Dans cette optique, notre groupe a développé une procédure d'induction standardisée qui inclut des stresseurs et des situations inductrices d'émotions "réalistes" et qui favorise l'enregistrement de parole spontanée ainsi que de phrases standards.

Dans cette approche, l'amélioration de la performance du système ASV est testée en comparant la performance du système avec un "entraînement structuré" - qui consiste à construire des modèles de locuteurs basés sur un ensemble d'enregistrements de parole obtenus dans les conditions destinées à induire des variations de l'état émotionnel - et avec un "entraînement neutre" dans lequel les modèles de locuteurs sont construits avec les enregistrements de parole neutre uniquement:

Une version préliminaire de cette procédure informatisée d'induction émotionnelle et d'enregistrement de parole a été utilisée dans le cadre d'une expérience pilote réalisée en collaboration avec plusieurs groupes de chercheurs européens (projet "Verivox", ESPRIT/BRA, [Kar98],

[1] Les recherches présentées ont été financées par un projet ESPRIT/BRA (VeriVox) dans la phase initiale, puis par un Fond National de Recherche Scientifique suisse (2151-49'685.96) et par le projet plurifacultaire "Prosodie" à l'Université de Genève.

[Sch98]). Les résultats de la comparaison de performance du système avec un "entraînement structuré" et avec un "entraînement neutre" sont brièvement présentés plus loin (v. point 3.1.).

2.2. Définir un ensemble de paramètres acoustiques plus optimaux

Le choix d'un ensemble de paramètres acoustiques optimaux pour la vérification automatique du locuteur doit satisfaire un certain nombre de critères. Nolan a proposé les 6 critères suivants [Nol83] : variabilité interlocuteur maximale, variabilité intra-locuteur minimale, résistance à l'imitation, disponibilité, stabilité lors de la transmission et facilité de mesure ; les deux premiers critères étant prioritaires.

Afin de définir les paramètres les plus adaptés à la vérification automatique du locuteur, nous avons entrepris d'enregistrer une grande base de données de parole comprenant un grand nombre de locuteurs et des variations vocales intra-locuteur liées à des variations contrôlées de l'état émotionnel des locuteurs, puis d'effectuer des analyses acoustiques visant à définir la variabilité inter et intra-locuteur pour un ensemble de paramètres mesurés. Cette recherche est décrite de manière plus détaillée dans la dernière section de cette article (4. Prolongements - Recherche en cours).

3. RÉSULTATS D'EXPÉRIENCES PILOTES

3.1. Première étude pilote

En collaboration avec plusieurs groupes de recherche dans le cadre d'un projet ESPRIT/BRA [Kar98], [Sch98], nous avons réalisé un premier test concernant la possibilité d'améliorer la performance d'un système ASV en incluant d'avantage de variabilité dans les modèles des locuteurs construits par le système.

Les changements induits au niveau de la parole dans le cadre de cette étude incluaient une série de modifications volontaires et involontaires de la parole (intensité, rapidité...) ainsi qu'une induction de stress psychologique (surcharge cognitive). 50 locuteurs suédois masculins ont utilisé le programme d'induction émotionnelle. La base de données de parole a été conçue de manière à contenir les enregistrements de parole nécessaires pour enrôler les locuteurs dans le système ASV avec un "entraînement neutre" et avec un "entraînement structuré" ainsi que le matériel nécessaire pour tester le système ASV avec différents styles de parole après l'entraînement. Tous les enregistrements ont été effectués lors d'une seule session pour chaque locuteur. Une version du système ASV du projet CAVE à été utilisée (v. [Bim97]). Lors du test du système ASV, la méthode d'entraînement qualifiée "d'entraînement structuré" a été comparée à la méthode standard dans laquelle les modèles de locuteurs sont construits avec les enregistrements de parole neutre uniquement.

Les résultats de cette étude pilote ont montré une diminution de L'EER de 32% (Equal Error Rate, pour un groupe de locuteur uniquement masculin) en passant de l'entraînement neutre (EER = 2.7%) à l'entraînement structuré (EER = 1.8%) avec des seuils indépendants du locuteur et de 41% avec des seuils dépendants du locuteur. L'avantage de l'entraînement structuré sur l'entraînement neutre est encore plus clair si l'on considère la réduction du taux de faux-rejets pour un taux donné de fausses-acceptations. Le taux de faux-rejets est réduit de 2.7% (EER) à 1.4% en passant à l'entraînement structuré, ce qui correspond à une réduction de 48% du taux de faux-rejets. (v. [Kar98] et [Sch98] pour de plus amples précisions).

Une première tentative d'évaluation de la variabilité intra-locuteur de différents paramètres liée aux différents styles de parole enregistrés dans cette étude a été effectuée par Nolan et collaborateurs à l'Université de Cambridge. Ils ont mesuré les durées segmentales et les bandes de fréquences des formants au centre des voyelles pour les enregistrements de 6 locuteurs issus de la base de données décrite ci-dessus. Bien que la quantité d'enregistrements analysés soit réduite les résultats indiquent, conformément aux attentes, que ces mesures sont affectées fortement et systématiquement par les variations intra-locuteurs. Ils ont, notamment, pu montrer, dans la condition de "parole rapide" et de "stress psychologique", un pattern de dépériphéralisation des voyelles mesurées par les bandes de fréquences du premier et du deuxième formants. (v. [Kar98] pour plus de détails).

3.2. Deuxième étude pilote

La base de données utilisée dans la deuxième étude pilote [Kam00] comportait trois phrases et des séquences de 10 chiffres prononcées par 10 locuteurs/acteurs allemands avec différents styles de parole (par exemple rapide, lente, forte, etc...) et avec des émotions simulées. Les locuteurs ont simulé des états émotionnels subtils, proches d'états qui pourraient survenir dans l'interaction avec des systèmes ASV. Un système ASV dépendant du texte – développé à la TU-Berlin avec le *Toolkit HMM* de HTK – a été utilisé. Dans cette étude le système a été entraîné uniquement selon la méthode standard (c'est à dire avec des enregistrements de parole neutre) sa performance a ensuite été testée avec les enregistrements incluant les différents styles de parole et la parole émotionnelle. Pour chaque locuteur, les enregistrements ont été effectués lors de trois sessions séparées par des intervalles de 4 jours (au minimum) à 6 semaines (au maximum).

Les résultats présentés ci-dessous (figures 1 et figure 2) comparent l'EER et la variabilité de plusieurs paramètres acoustiques pour différents styles de parole. 9 paramètres ont été choisis pour cette comparaison. Il s'agit des valeurs du 1er, du 2ème et du 3ème formant au milieu des phonèmes pour les voyelles /a/ et /i/ et pour la nasale /m/.

Les différences absolues en Hz entre les valeurs mesurées pour chaque formant et la moyenne des valeurs pour chaque formant dans les productions neutres, divisée par cette même valeur ont été additionnées pour chaque locuteur. Les figures 1 et 2 montrent un résultat typique : les conditions dans lesquelles une plus grande différence de valeur pour les paramètres mesurés relativement à la condition neutre apparaît correspondent aux conditions pour lesquelles l'EER est plus élevée. La corrélation est impressionnante, si l'on considère qu'il existe un grand nombre de paramètres autres que ceux que nous avons mesurés qui sont susceptibles d'avoir une influence sur l'EER.

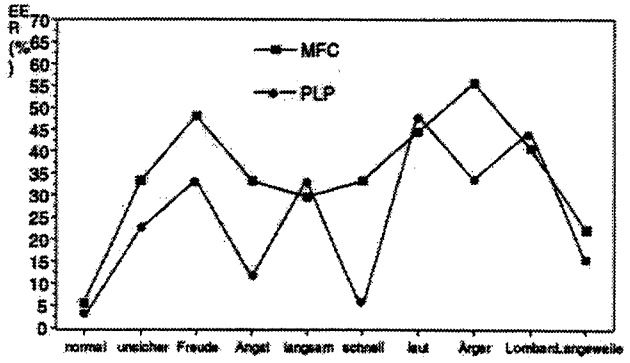


Figure 1: Valeurs d'EER pour différents styles de paroles

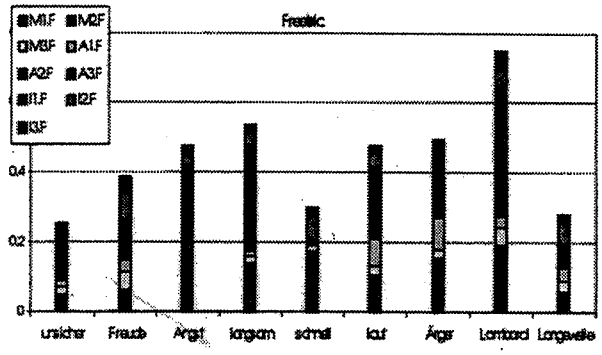


Figure 2: Total des différences de position des formants

Bien que les données utilisées ne comportaient que des modifications de parole liées à des "émotions" plutôt subtiles, les taux d'EER sont montés à plus de 50% lorsque l'entraînement du système n'est réalisé qu'avec les enregistrements de parole neutre et le système testé avec les enregistrements de différents styles de parole (y compris émotionnelle). Les paramètres que nous avons mesurés dans cette étude donnent une première indication concernant l'existence de changements spectraux spécifiques pour différents styles de parole.

4. PROLONGEMENTS - RECHERCHE EN COURS

Face à ces résultats préliminaires encourageants, nous avons initié une étude de plus grande envergure dans le but, d'une part de tester la possibilité d'améliorer la performance des systèmes ASV avec des enregistrements incluant d'avantage de modifications liées à l'état émotionnel des locuteurs, d'autre part d'examiner plus systématiquement quels paramètres acoustiques pourraient être avantageusement utilisés par les systèmes ASV.

4.1 Programme d'induction

Nous avons dans cette perspective développé un nouveau programme d'induction permettant de placer les locuteurs face à différentes tâches interactives destinées à modifier leur état émotionnel et à enregistrer des échantillons de parole émotionnelle spontanée et standard (lue). Nous avons inclus dans ce programme :

–Une tâche destinée à susciter un état de stress lié à une surcharge cognitive - induite en demandant aux locuteurs d'effectuer un test de logique verbale tout en réagissant à des stimulations sonores.

–Une tâche présentée comme un jeu informatisé de poursuite et de fuite permettant de susciter une forme de stress liée à l'activité de coordination visuo-motrice et aux bénéfices/pertes mis en jeu pour le locuteur.

–Une tâche de "discours public" ("public speech") - paradigme habituellement utilisé en psychologie comme inducteur d'anxiété.

–Une situation associant la frustration à l'injustice - produite par un mauvais fonctionnement (contrôlé) du programme associé à un feed-back négatif - destinée à induire de l'irritation chez les locuteurs

–Une combinaison de la méthode Velten [Vel68] avec des extraits musicaux afin d'amener les locuteurs à produire des expressions de "tristesse" d'une part et de "joie" d'autre part.

–A la fin de chaque session d'enregistrement, les locuteurs ont, en outre, lu plusieurs phrases standards avec la consigne de modifier leur expression vocale afin d'exprimer des émotions définies par des scénarios.

4.2 Enregistrement d'une base de données

A l'aide de ce programme nous avons enregistré 120 locuteurs (30 locuteurs allemands, 25 locuteurs anglais, 65 locuteurs français). L'utilisation, pour le test, d'un système ASV indépendant du texte nous a autorisé à ajouter des enregistrements de parole spontanée aux enregistrements constitués de phrases standards (lues).

4.3 Entraînement structuré d'un système ASV

En collaboration avec Ensigma, nous avons modifié puis prétesté un de leurs systèmes ASV afin de nous permettre de construire non seulement les modèles habituels pour les locuteurs ("entraînement neutre") mais également de réaliser un "entraînement structuré" du système à l'aide des enregistrements réalisés dans les contextes émotionnels. Nous allons prochainement, en collaboration avec Ensigma, utiliser ce système afin de comparer les performances de ces deux types d'entraînements réalisés à l'aide des données récoltées selon la procédure décrite ci-dessus.

4.4 Analyses acoustiques

Bien que l'analyse automatique soit certainement moins fiable que analyse manuelle, le grand nombre d'enregistrements effectués (100 à 120 par locuteur pour 120 locuteurs, soit environ 13'000 enregistrements) exige que les analyses acoustiques soient effectuées de manière automatique ou, à défaut, semi-automatique. Nous avons donc développé un ensemble de routines informatiques pour réaliser l'extraction automatique d'un grand nombre de paramètres acoustiques.

Afin de définir à quel degré différents paramètres acoustiques varient systématiquement avec différents styles de parole ou au contraire restent stables. Des paramètres différents de ceux habituellement utilisés dans les systèmes de vérification ont été choisi. Les considérations générales suivantes ont été prises en compte lors du développement des programmes pour l'analyse acoustique automatisée :
La sélection des caractéristiques acoustiques mesurées a été limitée à des caractéristiques qui peuvent être interprétées en terme de production de la parole et peuvent être facilement utilisées par les systèmes ASV actuels, de même que par des synthétiseurs de parole et par des systèmes de reconnaissance de la parole.
Les auditeurs naïfs perçoivent la qualité vocale comme une propriété générale de la parole orale. De nombreux chercheurs ont probablement fondé sur cette base leur conviction que la qualité vocale pouvait être décrite en utilisant des mesures moyennes calculées sur des phrases entières ou sur les parties voisées et non-voisées de phrases entières. Cependant, bien que les mesures moyennes suprasegmentales peuvent effectivement décrire certains aspects de la qualité vocale, une partie importante des informations contenues dans le signal vocal dynamique sont estompées par le calcul de valeurs moyennes à long terme. Les routines qui ont été développées dans le cadre de ce projet visent à extraire des paramètres qui permettent de décrire ces aspects dynamiques en détail.

Parmi les paramètres sélectionnés figurent la description du contour de F0, du contour de l'énergie et la structure des formants, ainsi que la distribution de l'énergie à l'intérieur de différentes bandes de fréquences pour des segments définis sur le plan phonétique.

Avant tout, la longueur totale des séquences de parole est calculée, une détection du signal et des pauses, puis une détection des parties voisées et non-voisées du signal sont effectuées. Un programme permettant de détecter des "syllabes" de manière relativement grossière en se basant sur la stylisation du contour d'énergie (v. ci-dessous) a également été défini. Les "voyelles accentuées" sont également détectées sur la base d'un ensemble de règles relativement simples. Comme pour les "syllabes", cette détection "grossière" est réalisée de manière automatique au prix d'une perte de précision. Une correction manuelle sera utilisée pour rectifier les erreurs de la détection automatique. Des contours d'énergie et de F0 sont également extraits automatiquement. Cette stylisation automatique a été choisie pour décrire les caractéristiques dynamiques de différents signaux, elle ne peut donc pas être interprétée en termes perceptifs. La stylisation des contours d'énergie est réalisée en filtrant le signal (filtre passe-bas à 30 Hz) puis en identifiant localement les minimas et les maximas "importants" ("l'importance" des maximas/minimas est définie sur la base d'un ensemble de règles complexes). Le contour de F0 est stylisé, en suivant également une procédure basée sur des règles. Finalement une stylisation du contour des formants à l'intérieur des syllabes accentuées est réalisée.

5. Conclusion, perspectives

L'analyse des données enregistrées à l'aide du programme d'induction émotionnelle décrit ci-dessus est en cours. Les premiers résultats obtenus ainsi que les résultats d'études pilotes rapportés dans ce qui précède nous confortent dans l'idée que la vérification automatique du locuteur gagnerait à baser ses modèles des locuteurs sur une connaissance plus approfondie de la manière dont les modifications de l'état interne des locuteurs affectent la réalisation acoustique de leurs expressions vocales. Les analyses et les tests qui seront effectués dans le cadre de la recherche que nous avons initiée devraient permettre de considérablement progresser dans cette direction.

Bibliographie

[Bim97] Bimbot F., Hutter H.P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J. (1997), "Speaker Verification in the Telephone Network: Research activities in the CAVE Project", proc. EUROSPEECH'97, Vol. 2, pp. 971-974.

[Kam00] Kamceva, T. (2000), "Formantverläufe bei verschiedenen Sprechweisen und deren Zusammenhang mit der Fehlerrate eines textabhängigen Sprecherverifizierungssystem", Unpublished Diploma thesis, TU-Berlin.

[Kar98] Karlsson I., Bänziger T., Dankovicova J., Johnstone T., Lindberg J., Melin H. Nolan F., Scherer K. (1998), "Within-speaker variability due to speaking manners", proc. ICSLP'98, pp. 207-210.

[Nol83] Nolan F. (1983), The phonetic bases of speaker recognition., Cambridge University Press.

[Sch98] Scherer K., Johnstone T., Bänziger T. (1998), "Automatic verification of emotionally stressed speakers : The problem of individual differences", proc. SPECOM'98, pp. 233-238.

[Vel68] Velten, E. (1968), "A laboratory task for induction of mood states", Behavior Research and Therapy, 6, pp. 473-482.

Reconnaissance Automatique du Locuteur en Milieu Bruité -Cas de la SOSM-

H. SAYOUD, S. OUAMOUR, N. KERNOUAT et M.K. SELMANE*

*USTHB, Institut d'Electronique, BP 32 Bab Ezzouar, Alger, Algérie
*Email: sayoud@ifrance.com

Abstract

In this paper we are interested by the robustness of a recent method used in automatic speaker recognition. It is called the SOSM (Second Order Statistical Measure) using the gaussian measure of likelihood **µG** and the covariance measure of likelihood **µGc**. Tests of recognition are done in a population of 37 speakers extracted from the database of TIMIT and according to a text independent identification. Three types of noise are used: the gaussian white noise, the noise of rumpus and the noise of car; with several SNR, namely: 96, 18, 12, 6 and 0 dB. The Mel-spectral resolution is fixed to 36 Mel coefficients per 8 kHz, for an optimal identification performance (results of a previous work proposed for publication).
The results exposed in our figures show that the score obtained by the **µGc** measure is better than the score obtained by the **µG** measure for the case of the normal voice; on the other hand for the case of the telephonic band (called also FTIMIT) it seems that the **µG** gives better results than the **µGc**.
Concerning the type of noise used, we notice that the noise of car doesn't perturb severely the speaker recognition, contrarily to the other types of noise. And globally, the rates of recognition obtained reach the score of 100% in a non noised background but fall quickly at 6 dB.

Mots clés

Reconnaissance du locuteur, robustesse, SOSM.

1. Introduction

Plusieurs études effectuées par Bimbot [Bim95], Magrin-Chagnolleau [Mag95] et Bonastre [Bon97] ont montré l'efficacité d'une méthode récente, basée sur la mesure de vraisemblance gaussienne symétrique, en identification du locuteur. Des taux de reconnaissance proches de 100% pour de larges populations ont été, ainsi, obtenus. Cependant le fait d'obtenir de bons scores en chambre sourde (96 dB de RSB) n'est pas concluant du fait qu'en pratique nous parlons toujours en présence de bruits ambiants (chahut, bruit de moteur, diaphonie , etc...), voilà pourquoi nous avons entrepris une étude sévère en milieu bruité par les trois types de bruit qui sont rencontrés le plus souvent (chahut, bruit blanc gaussien, et bruit de voiture) et pondérés à des RSB différents. Notre intérêt s'est porté aussi vers une comparaison entre les deux mesures de vraisemblance **µG** et **µGc** [Bim95]. Notre but est, essentiellement, de donner une indication qualitative et quantitative quant à l'utilisation de la SOSM (Second Order Statistical Measures) dans le cas de la parole corrompue soit par un bruit ambiant ou par un filtrage du type téléphonique, pour conseiller les utilisateurs de cette technique en milieu réel (milieu industriel) sur la mesure à employer.
Dans ce qui suit, nous traitons l'identification du locuteur indépendante du texte, dans un ensemble fermé et dans un environnement corrompu.

2. Base de Données des Locuteurs

La base de données des locuteurs est extraite de la base TIMIT [Fis86]. Le nombre de locuteurs est de 37 soit 15 féminins et 22 masculins, avec un accent nord-américain. La durée approximative des phrases est de 9 s pour l'apprentissage et de 7 s pour le test. Les enregistrements sont faits avec un microphone de haute qualité, sur 16 bits et avec une fréquence d'échantillonnage de 16 kHz.

3. Pré-Traitement

Le pré-traitement est constitué des étapes d'analyse du signal de parole précédant l'apprentissage et le test d'identification. Ainsi, chaque phrase est analysée selon les étapes suivantes.

- Détection et suppression des zones de silence
- Evaluation du spectre d'énergie par une transformée de Fourrier rapide.
- Le spectre d'énergie passe ensuite à travers un banc de filtre de 36 canaux étendus sur une bande de 8 kHz.
- Un écrétage à seuil d'énergie [Dau83] est utilisé pour limiter le rang dynamique du signal.
- Finalement les coefficients d'énergie (ou énergies Mel) issus du banc de filtre pour chaque fenêtre sont stockés dans des vecteurs appelés les MFSC ou *Mel Frequency Spectral Coefficients*. Ce sont les caractéristiques du locuteurs utilisées dans cette étude.

4. Méthode utilisée pour la reconnaissance du locuteur : la "SOSM"

Cette méthode a été introduite par F. Bimbot, I. Magrin-Chagnolleau et L. Mathan en 1995, elle est basée sur une métrique gaussienne du type µG ou µGc [Bim95].

Dans cette méthode on exploite le fait que les caractéristiques spectrales de tout locuteur, pour une phrase longue, suivent une loi gaussienne stationnaire au second ordre. Les étapes de cette méthode sont les suivantes.
D'abord, on procède à l'extraction des MFSC (ou Mel énergies), puis pour chaque prononciation on fabrique le vecteur moyenne x et la matrice de covariance X ; ainsi il existe une moyenne x pour chaque locuteur et une covariance X pour chaque locuteur. Le couple (x,X) représente la référence statistique d'ordre 2 pour le locuteur $\mathcal{X}$ [Bim95] utilisé dans le dictionnaire des références.
En phase de test, une modélisation similaire de la phrase de test générera le couple (y,Y) représentant le modèle statistique de test pour le locuteur inconnu $\mathcal{Y}$.
Le test d'identification est basé, alors, sur la distance minimale (plus proche voisin) au sens de la métrique statistique du 2e ordre: **µG** ou **µGc**. Dans nos expériences nous avons utilisé les variantes symétriques supposant une durée d'apprentissage proche de celle du test, en l'occurrence: la $\mu G_{0.5}$ appelée mesure de vraissemblance gaussienne symétrique et la $\mu Gc_{0.5}$ appelée mesure de vraissemblance gaussienne symétrique à covariance [Bim95].
La première distance est définie par

$$\mu_{G0.5}(X,Y)=\frac{1}{2}\left[\alpha+\frac{1}{p}\delta^T\left(X^{-1}+Y^{-1}\right)\delta\right]-1 \quad (1)$$

avec

$$\alpha=\frac{1}{p}\left(tr\left(YX^{-1}\right)+tr\left(XY^{-1}\right)\right) \quad (2)$$

$$\delta=y-x \quad (3)$$

p étant la dimension du vecteur des caractéristiques acoustiques et "tr" dénote la trace d'une matrice.
$\mathcal{X}$ représente la référence et $\mathcal{Y}$ représente le locuteur à reconnaître.
De même la mesure $\mu Gc_{0.5}$, qui est une variante de la mesure précédente, est définie par :

$$\mu_{Gc0.5}(X,Y)=\frac{1}{2}\alpha-1 \quad (4)$$

Les deux distances seront testées simultanément pour évaluer leurs robustesses comparatives en milieu corrompu.

5. Résolution spectrale optimale

Une importante question pouvant se poser lors du choix de la dimension optimale des caractéristiques acoustiques du locuteur est la suivante :
Existe-il un rapport entre la résolution spectrale et la fiabilité de modélisation des paramètres acoustiques et prosodiques du signal de parole ?
Si ce rapport (intuitif) existe, alors une deuxième question se pose : quel est le nombre optimal de filtres Mel pour cette tâche ?

Ce problème nous a incité à tester différentes résolutions spectrales allant du modèle classique à 12 filtres Mel jusqu'au modèle complexe à 60 filtres Mel. Les résultats de cette étude antérieure (en cours de publication) ont montré que la résolution Mel-spectrale à 36 filtres Mel sur 8 kHz, impliquant 36 coefficients MFSC, est la résolution optimale en reconnaissance du locuteur, si on ne normalise pas ces coefficients MFSC (par rapport à l'énergie moyenne). Voilà pourquoi nous avons choisi, pour cette étude, une résolution de 36 canaux dans le banc de filtres.

6. Tests d'identification dans le cas non bruité

L'étude suivante consiste à identifier les 37 locuteurs de TIMIT [Fis86] par la méthode SOSM [Mag95], en milieu non bruité. Deux cas sont prévus : L'identification sur la bande [0-8 kHz] et l'identification sur la bande téléphonique [300-3400 Hz] notée FTIMIT. La dimension des MFSC est choisie égale à 36 coefficients, répartis sur la bande spectrale [0-8 kHz].

6.1. Résultats de cette étude

Deux constats intéressants sont à noter :

- Pour la bande [0-8000 Hz], le taux d'identifications est de 100% avec la **µG** et la **µGc** ce qui implique une identification fiable en milieu non bruité.
- Pour la bande [300-3400 Hz] notée FTIMIT, le taux d'identifications est de 97,29% pour la **µG** et de 94,59% pour la **µGc**. Ainsi la **µG** semble plus précise que la **µGc** sur la bande téléphonique.

6.2. Discussion et conclusion

En milieu sain les deux mesures **µG** et **µGc** sont très fiables pour la reconnaissance du locuteur, toutefois sur la bande téléphonique la **µG** agit mieux que la **µGc**, à cause des informations supplémentaires contenues dans le vecteur moyenne des caractéristiques acoustiques et qui permettent de mieux distinguer les locuteurs.

7. Tests d'identification en milieu bruité

La deuxième étude consiste à identifier les 37 locuteurs de TIMIT par la méthode SOSM, en milieu bruité. Les bruits utilisés sont décrits ci-dessous, avec une pondération variable allant de 0 dB à 18 dB. Deux cas sont prévus: l'identification sur la bande [0-8000 Hz] et l'identification sur la bande réduite [300-3400 Hz] notée FTIMIT.

7.1. Bruits utilisés

Nous avons utilisé trois types de bruits pour essayer de diversifier le milieu environnant. Ce sont :

- le bruit blanc gaussien noté **BBG**,
- le bruit de **chahut** ou de foule obtenu après mixage de plusieurs signaux de dialogues humains,

- le bruit de **voiture** obtenu après mixage de plusieurs bruits de moteur causés par le passage de véhicules dans une route à moyenne circulation.

Le bruitage est pondéré par le RSB choisi, allant de 0 dB jusqu'à 18 dB. Notons que le signal original non bruité sera symbolisé par un RSB de 24 dB pour simplifier la représentation graphique des courbes de résultat.

7.3. Discussion et conclusion

Il s'avère que les distances **µG** et **µGc** n'agissent pas de la même manière en milieu bruité. Ainsi sur la bande [0-8 kHz] (figures 1 et 2) la **µGc** est plus robuste à cause de l'élimination des informations du bruit contenues dans le vecteur moyenne des caractéristiques ; tandis que sur la bande [300-3400 Hz] (figures 3 et 4) la **µG** est nettement meilleure ; ce qui favorise son utilisation dans les applications téléphoniques.
De plus nous remarquons, surtout sur la figure 2, que le bruit de voiture n'altère pas la qualité de la reconnaissance contrairement à ce qu'on pouvait penser vu le caractère agressif de ce dernier au niveau de l'ouïe humaine. Cette remarque encourage l'utilisation des systèmes d'identification dans des endroits à proximité des routes et des autoroutes.
D'une manière générale la SOSM est assez robuste vis-à-vis du bruit ambiant additif, mais ce qui est marquant est la défaillance observée dès que l'on passe en dessous de 6 dB de RSB.

8. Synthèse générale

Nous pouvons déduire, des observations précédentes, qu'avant de concevoir un système pour la reconnaissance du locuteur, il faut d'abord faire une étude du milieu ambiant : RSB moyen, bande passante du canal, type de bruits susceptible de bruiter le signal de parole etc.
Ceci dans le but de choisir la meilleure méthode d'identification et le type de distance appropriée (cas de la SOSM) et pour assurer une qualité de reconnaissance optimale.
Les tests faits en milieu bruité (BBG, chahut et bruit de voiture) ont montré que le codage MFSC à 36 Mels apporte une grande quantité d'informations propre au locuteur et aide à mieux le distinguer en milieu bruité. Ainsi la dimension de 36 coefficients par 8 kHz (à l'échelle Mel) apparaît comme une dimension favorable pour les systèmes robustes de reconnaissance du locuteur.
De même la technique SOSM utilisant la mesure **µGc** semble aussi très robuste vis-à-vis des bruits additifs utilisés. Cependant, sur la bande [300-3400 Hz] la distance optimale est la **µG**, que ce soit en milieu sain ou en milieu bruité.
Les résultats obtenus pour le bruit de voiture prouvent ainsi que ce bruit n'est pas gênant en reconnaissance du locuteur, contrairement à ce qu'on pouvait penser vu la gêne auditive générée par ce dernier. Par conséquent, un système de vérification de locuteur peut être aisément implanté dans les endroits à proximité des grandes routes (station-service, autoroute, station de bus etc.). Alors que le chahut, si bien filtré par notre système nerveux, s'avère être plus gênant pour identifier un locuteur. Il s'en suit qu'avant toute procédure de vérification de locuteur, il faut s'assurer que la zone d'enregistrement microphonique ne soit pas une zone à chahut ou une zone riche en bruit de foule (marché, bureau de poste, gare etc.).
Finalement, pour plus de rigueur, nous tenons à faire remarquer que les résultats de ce travail de recherche sont typiques à une seule méthode de reconnaissance du locuteur, qui est la SOSM.

Références bibliographiques

[Bim95] F. BIMBOT, I. MAGRIN-CHAGNOLLEAU, et L. MATHAN 1995, "Second-Order Statistical measures for text-independent Broadcaster Identification". Speech Communication, Volume. 17, Number, 1-2, August 1995, pp. 177-192.

[Bon97] F. BONASTRE et L. BESACIER 1997, "Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur". Actes du 4ème Congrès Français d'Acoustique, pp 357-360, Marseille 14-18 April 1997.

[Dau83] B.A. DAUTRICH, L.R. RABINER and T.B. MARTIN 1983, "The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer", Bell System Technical Journal, 1983.

[Dod98] G. R. DODDINGTON 1998, " Speaker Recognition Evaluation Methodology. An Overview and Perspectives", RLA2C Avignon, 20-23 April 1998, pp 60-66.

[Fis86] W. FISHER, V. ZUE , J. BERNSTEIN and D. PALLET 1986, "An acoustic-phontic database", JASA, suppl. A, Vol. 81(S92) 1986.

[Mag95] I. MAGRIN-CHAGNOLLEAU, J. F. BONASTRE et F. BIMBOT 1995, "Effect of Utterance Duration and phonetic Content on Speaker identification Using Second-Order Statistical Methods", ESCA EUROSPEECH'95, vol. 1, pp 337-340, sep. 1995, Madrid.

[Rey94] D.A. REYNOLDS 1994, "Speaker identification and verification using Gaussian Mixture speaker models", Workshop on Automatic Speaker Recognition, identification and verification", April 1994, Martigny, Switzerland, pp. 27-30.

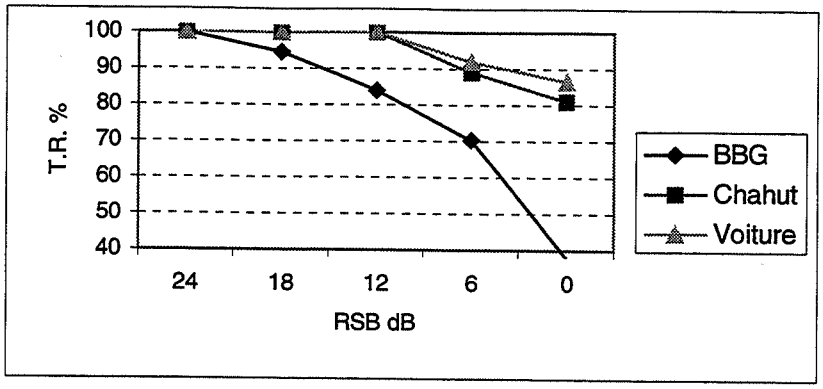


Figure 1 Taux de reconnaissance obtenus, après différents bruitages, sur TIMIT par la μG

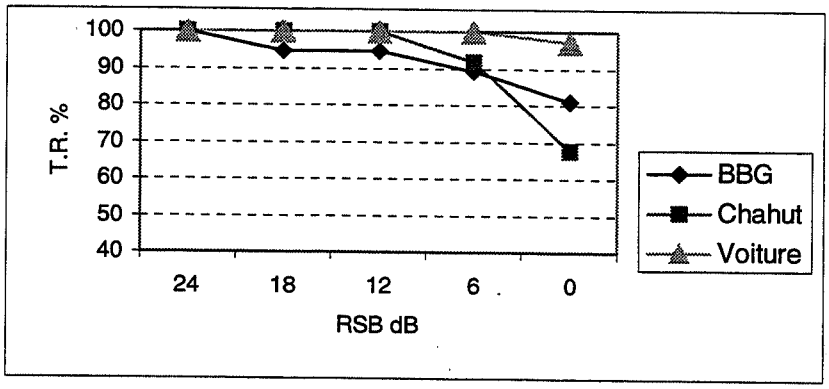


Figure 2 Taux de reconnaissance obtenus, après différents bruitages, sur TIMIT par la μGc

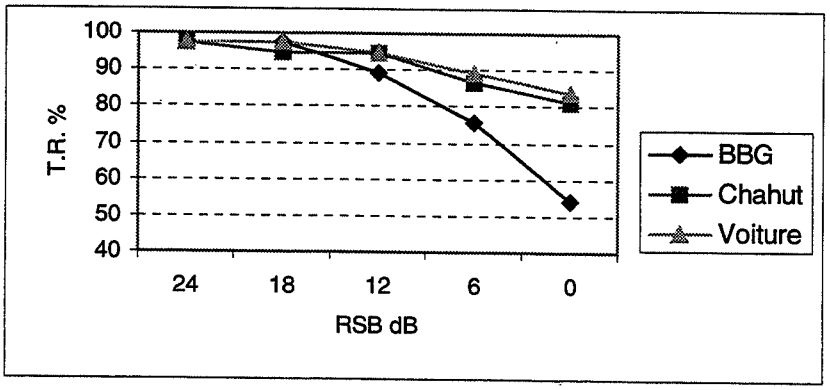


Figure 3 Taux de reconnaissance obtenus, après différents bruitages, sur FTIMIT par la μG

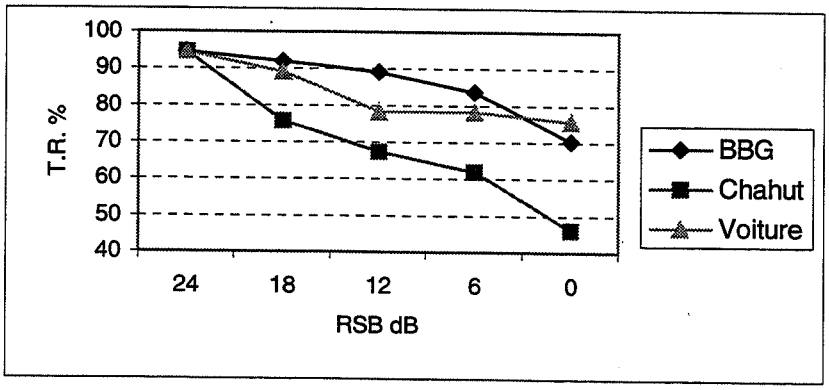


Figure 4 Taux de reconnaissance obtenus, après différents bruitages, sur FTIMIT par la μGc

Adaptation robuste de modèles HMM pour la vérification du locuteur dépendante du texte

Johnny Mariéthoz *, *Frédéric Bimbot* **

* IDIAP - BP 592, CH-1920 Martigny, Suisse
** IRISA (CNRS & INRIA) - Campus de Beaulieu, 35042 Rennes cedex, France

ABSTRACT

When deploying a secure system based on speaker verification, the limited amount of training data is usually critical. Indeed, the enrollment procedure must be fast and user-friendly. An incremental training of HMM speaker models, based on a MAP (Maximum A Posteriori) adaptation technique is used in order to make the enrollment more robust with only one or two utterances of the client password. This paper presents the improvements which can be achieved, in term of verification performance and stability of the decision thresholds. Our results highlight the benefits of MAP adaptation in conjunction with a synchronous alignment approach.

1. INTRODUCTION

La vérification du locuteur suscite un intérêt croissant de la part des fournisseurs de services téléphoniques, dans la mesure où ces techniques permettent de mieux sécuriser les transactions vocales sur les différents réseaux de télécommunications, en offrant la possibilité de réduire les risques de fraude sans nécessiter l'implantation d'équipement supplémentaire chez l'abonné. Cependant, des difficultés spécifiques existent pour ce type d'applications commerciales, une d'entre elle étant la nécessité de garantir une mise en oeuvre rapide du service pour tout nouvel utilisateur. En pratique, les applications visées doivent être opérationnelles à partir d'une ou deux sessions d'entraînement, ce qui limite considérablement la représentativité des données d'apprentissage, que ce soit en termes de couverture de la variabilité individuelle au cours du temps ou de type de microphone et de canal de transmission observé.

Pour remédier à ce problème, une solution consiste à affiner, au fil de l'utilisation du système, les modèles caractéristiques de chaque client avec les énoncés produits par ce client à l'occasion d'utilisation précédentes du service, afin d'acquérir progressivement des données plus représentatives des différentes conditions d'utilisation de l'application par ce client.

Les travaux présentés dans cet article se placent dans le contexte d'un formalisme probabiliste du problème de la vérification du locuteur, où la décision est prise à partir d'un rapport de vraisemblance fourni par le modèle spécifique du client et un modèle indépendant du locuteur (appelé *modèle du monde*).

Nous utilisons les techniques d'adaptation Bayésienne pour effectuer l'apprentissage incrémental du modèle du client. Nous comparons tout d'abord l'impact, sur les performances du système, d'une approche incrémentale par adaptation à partir des données nouvelles par rapport à une approche par réapprentissage complet utilisant l'ensemble des données produites. Nous commentons ensuite nos observations sur la dérive des seuils de décision optimaux et nous présentons une solution permettant de remédier aux problèmes rencontrés, en utilisant également une technique Bayésienne pour estimer le modèle client initial. Enfin, nous mettons en évidence un avantage supplémentaire à utiliser une technique d'alignement synchrone pour calculer le rapport de vraisemblance sur une séquence d'états commune aux modèles du client et du monde.

Les travaux rapportés dans cet article sont effectués dans le contexte du projet Européen Telematics PICASSO [B+99] (Work-Package 5). Les expériences ont été réalisées avec la plate-forme logicielle commune *Picassoft* sur la base de données PolyVar / suisse romand selon un protocole expérimental défini par l'ensemble des partenaires.

2. CADRE GÉNÉRAL

2.1. Modèle probabiliste

L'approche utilisée dans l'ensemble de cet article s'appuie sur un formalisme probabiliste du problème de la vérification. Pour un enoncé de test noté Y prononcé par un locuteur proclamant l'identité X, on calcule le logarithme du rapport de vraisemblance:

$$s_X(Y) = log\left(\frac{\hat{\mathcal{P}}(Y|X)}{\hat{\mathcal{P}}(Y|\bar{X})}\right)$$

où $\hat{\mathcal{P}}(Y|X)$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par le locuteur proclamé et où $\hat{\mathcal{P}}(Y|\bar{X})$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par un autre locuteur.

Le modèle probabiliste correspondant à X (dit *modèle client*) est estimé à partir de données d'apprentissage composées d'énoncés prononcés par X. Le modèle correspondant à $\bar{X}$ (dit *modèle non-client*) est obtenu à partir d'énoncés semblables prononcés par d'autres locuteurs. Quand le modèle du non-client est le même pour tous les clients, ce qui est le cas dans ces travaux, on le désigne par *modèle du monde* (noté Ω).

Dans les travaux décrits ici, la vérification s'effectue sur un mot (ou un groupe de mot) issu d'un vocabulaire de 17 mots différents, ce vocabulaire étant commun à tous les clients. Les modèles probabilistes utilisés sont des HMM (Modèles de Markov Cachés) à topologie gauche-droite (un par mot) dont les fonctions

d'émission des états sont des mélanges de distributions Gaussiennes. Une spécificité importante de nos HMM réside en ce que les modèles client et le modèle du monde ont une topologie identique.

2.2. *Décision et types d'erreurs*

Dans les applications où il s'agit de prendre une décision binaire d'acceptation ou de rejet de l'identité proclamée, le score s_x est comparé à un seuil de décision choisi de façon à optimiser les performances du système dans une condition de fonctionnement particulière. Cette condition de fonctionnement est spécifiée par le rapport des coûts associés aux deux types d'erreur possibles: faux rejet, si un client authentique est rejeté par le système et fausse acceptation si un imposteur n'est pas détecté.

2.3. *Mesure des performances*

Les performances des approches décrites dans cet article sont présentées sous deux formes:

– une courbe DET [MP97] qui indique les caractéristiques du système en terme de pouvoir de séparation des clients et des imposteurs: plus la courbe DET est proche de l'origine, meilleure est la séparation apportée est le système.

– les performances du système dans une condition de fonctionnement équicoût, c'est-à-dire pour laquelle les deux types d'erreur sont considérées comme étant de gravité égale. Dans ce cas, la décision optimale vise à minimiser le Demi Taux d'Erreur Total (DTET), c'est-à-dire la moyenne arithmétique du taux de faux rejets et du taux de fausses acceptations. Les résultats présentés sont obtenus par réglage des seuils a posteriori c'est-à-dire en optimisant le DTET sur l'ensemble de test. Notons que, pour la condition de fonctionnement équicoût, le seuil Bayésien théorique sur le *logarithme* du rapport de vraisemblance est égal à 0.

3. Apprentissage incrémental

3.1. *Modalités d'apprentissage*

Une partie de notre étude consiste à comparer les performances du système selon deux modalités d'apprentissage des modèles du client, désignées par mode *batch* et mode *incrémental*.

Dans les deux modalités, un modèle client initial est estimé à partir de 1 répétition, provenant des 2ères sessions d'enregistrement, soit 2 énoncés au total. Le modèle initial ainsi obtenu sera désigné dans la suite par l'abréviation '12'. Précisons que l'algorithme d'apprentissage utilisé est l'algorithme des k-moyennes segmentales, c'est-à-dire un algorithme EM avec segmentation par Viterbi, où le modèle initial est le modèle du monde.

Dans le mode *batch*, on réestime complètement, après chaque nouvelle session, le modèle client à partir des données d'initialisation (2ères sessions) auxquelles on adjoint successivement la répétition des sessions ultérieures (session 3, puis 4, puis 5), soit des ensembles d'apprentissage constitués respectivement de 3, 4 et 5 énoncés. Chaque réestimation nécessite que soit conservés en mémoire les énoncés représentés sous forme acoustique (paramétrée). On désignera ces configurations par les abréviations 123, 1234 et 12345.

Dans le mode *incrémental*, on fait l'hypothèse que l'on a plus accès aux données acoustiques des sessions passées et que l'on doit se limiter à adapter le modèle client courant à partir du seul énoncé de la session précédente. Cette contrainte est imposée par la préoccupation de minimiser et de contrôler le volume occupé par les informations nécessaires à caractériser le client. On désignera par 12+3, 12+3+4 et 12+3+4+5 les configurations correspondant à l'apprentissage incrémental avec les sessions 3, 4 et 5 respectivement.

3.2. *Adaptation Bayésienne*

Les techniques d'adaptation Bayésienne sont couramment utilisées pour l'estimation statistique des modèles probabilistes utilisés en reconnaissance de la parole [GL94] car elles offrent un cadre théorique et une bonne efficacité pratique pour traiter des différents problèmes d'adaptation que l'on peut rencontrer dans les contextes applicatifs, que ce soit l'adaptation au locuteur, au canal, à l'environnement d'utilisation, etc (voir par exemple [MC99]).

Nous adoptons cette même approche pour l'apprentissage incrémental: nous considérons que le problème posé revient à adapter le modèle client estimé sur les sessions initiales, à partir de nouvelles observations (en l'occurrence les données client provenant des sessions ultérieures) [Mok98]. Par ailleurs, on se place dans le cadre de l'adaptation supervisée, c'est-à-dire sous l'hypothèse que les données servant à adapter le modèle proviennent effectivement du client. Des travaux parallèles [F+00] étudient le comportement de l'apprentissage incrémental en cas d'attaques d'imposteurs.

En pratique, nous utilisons une version simplifiée de l'apprentissage Bayésien qui consiste à n'actualiser que les moyennes des distributions gaussiennes, selon la formule d'adaptation:

$$\mu_{n+1} = \frac{\alpha_n \mu_n + a m}{\alpha_n + a}$$

où μ_n et μ_{n+1} désignent respectivement les moyennes du modèle avant et après adaptation et où m représente la moyenne des données observées. Le poids α_n est pris égal au nombre de données utilisées pour estimer la valeur de μ_n et a correspond au nombre de valeurs observées pour calculer m. A chaque incrément, α est remis à jour: $\alpha_{n+1} = \alpha_n + a$.

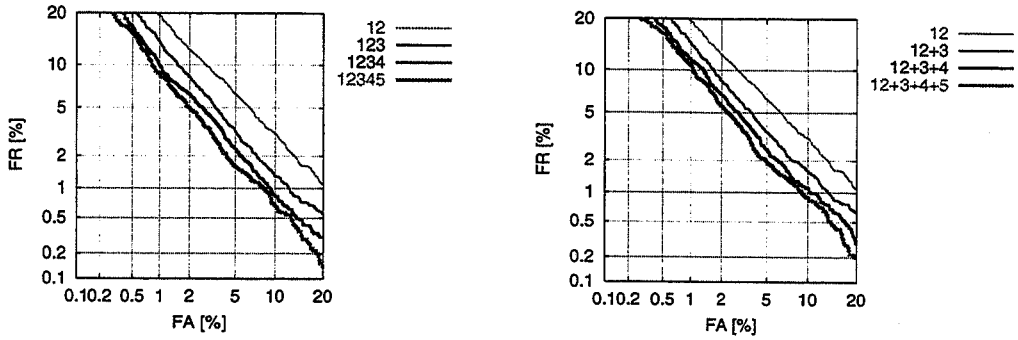
4. Performances

4.1. *Protocole d'évaluation*

La base de données utilisée comporte 17 mots de commande[1] provenant de la base de données PolyVar / suisse romand. La population des clients est constituée de 19 locuteurs (12 hommes et 7 femmes). Une autre population de 56 locuteurs (28 hommes et 28 femmes) est utilisées pour estimer le modèle du monde (56 énoncés). Les résultats expérimentaux sont obtenus à partir d'environ 6000 accès clients (soit, en moyenne, de l'ordre de 15 accès par client et par mot) et d'à peu près 12000 accès imposteurs (issus de la même population que celle des clients).

Les coefficients LPCC d'ordre 16, ainsi que les deltas et les delta-deltas sont utilisés pour la paramétrisation acoustique des énoncés. La topologie des modèles HMM des clients et du monde est identique, à savoir 2 états par phonème et 1 gaussienne par état.

1. annulation, casino, cinéma, concert, corso, exposition, galerie du Manoir, Gianadda, guide, Louis Moret, Manifestation, message, mode d'emploi, musée, précédent, quitter, suivant



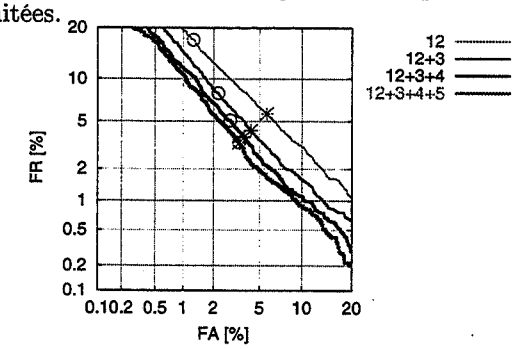
(a) batch (b) incrémental

Sessions	12	12+3	12+3+4	12+3+4+5
DTET [%] batch	5.67	4.07	3.61	3.12
DTET [%] incremental	5.67	4.26	3.73	3.39

FIG. 1 – *batch vs incrémental*

4.2. Résultats

La figure 1 présente, sous forme de courbes DET, les performances des deux protocoles d'apprentissage (incrémental vs batch). Ces figures mettent en évidence un avantage relativement marginal de l'approche batch. On retiendra donc que l'approche incrémentale ne semble pas dégrader les performances de façon sensible et qu'il est donc judicieux de l'utiliser dès lors que les capacités de stockage pour chaque client sont limitées.



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	-0.678	-0.353	-0.201	-0.092

FIG. 2 – *Dérive des seuils en mode incrémental*

En revanche, on observe, pour les deux méthodes, une dérive du seuil optimal pour le DTET en fonction du nombre de sessions prises en compte. Ceci est mis en évidence, pour l'apprentissage incrémental, sur la figure 2 où l'on peut comparer, les points de fonctionnements optimaux pour chaque configuration (représentés pas des croix) et le point de fonctionnement correspondant à un seuil fixe, optimisé sur la configuration à 5 sessions (représentés par un rond). Le même phénomène s'observe dans le cas de l'apprentissage batch. Cette dérive est génante, car elle rend nécessaire une estimation de seuil différente pour chaque réestimation ou chaque adaptation.

5. DÉRIVE DES SEUILS

5.1. Analyse diagnostique

Une analyse fine du comportement du système sur la tâche traitée indique que la dérive des seuils provient de la variation dans la qualité de l'estimation des modèles du client, estimation d'autant plus mauvaise que le volume de données utilisé est limité. En effet, en cas de données insuffisantes, les estimateurs $\hat{\mathcal{P}}(Y|X)$ des distributions clients $\mathcal{P}(Y|X)$ sont de très mauvaise qualité et induisent de ce fait un biais négatif dans la valeur du seuil optimal, par rapport au seuil Bayésien théorique (égal à 0 pour le DTET).

D'un point de vue pratique, ces mauvaises estimations se manifestent à deux niveaux: d'une part une mauvaise estimation des moyennes des gaussiennes dans les états du HMM client. D'autre part, un chemin de décodage inadéquat lors de l'alignement de l'énoncé de test avec le modèle client. Les scores de vraisemblance de chaque trame de test sont donc doublement entâchés d'erreur.

5.2. Apprentissage par adaptation

L'approche utilisée dans les expériences précédentes repose sur un apprentissage du modèle client initial (configuration 12) à partir des données d'entraînement correspondantes, en utilisant le modèle Ω comme initialisation de l'algorithme EM. Néanmoins, au cours des itérations, certains états peuvent devenir faiblement occupés, voire totalement désertés par les données d'apprentissage, ce qui a une influence néfaste tant sur les capacités de généralisation du modèle que sur la qualité de l'alignement qu'il peut fournir sur de nouvelles observations.

C'est pourquoi nous avons opté dans notre contexte de vérification *dépendante* du texte pour une approche d'estimation des modèles clients initiaux basée sur l'adaptation Bayésienne du modèle du monde, s'appuyant sur des résultats montrant l'intérêt de procéder ainsi en vérification du locuteur *indépendante* du texte [Rey97]. Seule l'intialisation du modèle client est modifée (étape 12), ensuite l'apprentissage reste identique (étape 12+3, etc).

En utilisant le même formalisme d'adaptation que précédemment, l'approche proposée revient donc à estimer les moyennes des gaussiennes des fonctions d'émission pour les HMM clients sous la forme:

$$\mu_X = \frac{\beta \mu_\Omega + b m_X}{\beta + b}$$

où μ_X est la moyenne de la gaussienne du modèle client adapté, μ_Ω la moyenne de la gaussienne correspondante dans le modèle du monde, m_X la moyenne des données client associées à la gaussienne, et β et b les poids attribués au modèle du monde et aux données client respectivement.

Contrairement au cas précédent, les poids β et b ne peuvent être choisis égaux au nombre d'observations associées à la gaussienne considérée dans le modèle du monde et le modèle client respectivement, car, en pratique, $\beta \gg b$. C'est pourquoi on réécrit l'équation

précédente sous la forme:

$$\mu_X = \gamma \mu_\Omega + (1-\gamma) m_X$$

et l'on choisit une valeur de γ commune à toutes les gaussiennes de façon à optimiser les performances du système. Le paramètre γ correspond alors au poids relatif apparent du modèle du monde dans le processus d'adaptation. Dans nos expériences, nous avons testé les valeurs de γ égales à 0, 1/2, 2/3 et 3/4.

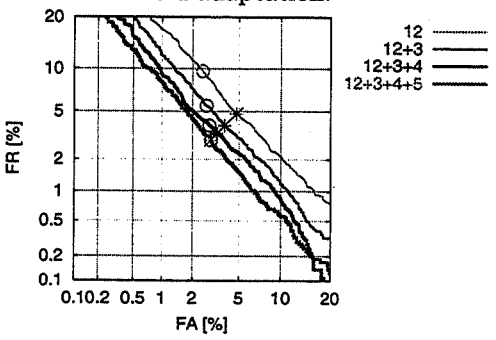
5.3. *Résultats*

La table 1 montre l'influence des valeurs de γ sur le DTET et sur la valeur du seuil optimal pour le modèle intial et ceux issus d'un apprentissage incrémental.

TAB. 1 – *Influence de γ sur le modèle initial et sur les modèles obtenus par apprentissage incrémental*

		γ = 0	1/2	2/3	3/4
12	seuil	−0.818	−0.252	0.053	0.190
	DTET[%]	6.26	4.76	4.85	5.22
12+3	seuil	−0.549	−0.146	0.103	0.239
	DTET[%]	4.15	3.69	3.88	4.15
12+3+4	seuil	−0.371	−0.068	0.059	0.276
	DTET[%]	3.40	3.24	2.88	3.49
12+3+4+5	seuil	−0.238	0.001	0.177	0.290
	DTET[%]	2.96	2.88	2.87	3.16

On observe que c'est pour les valeurs de γ de 1/2 et de 2/3 que les performances optimales sont obtenues avec une dérive des seuils moindre dans le second cas. Notons que dans ce dernier cas, le seuil optimal ne diffère du seuil théorique que de 5 à 15 % selon le nombre de sessions d'adaptation.



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	0.072	0.131	0.170	0.208

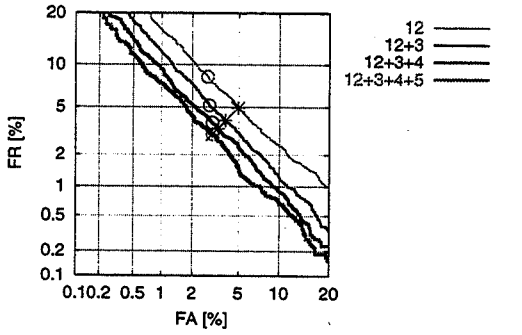
FIG. 3 – *Dérive des seuils avec une adaptation Bayesienne du modèle du monde avec $\gamma = 2/3$.*

5.4. *Alignement synchrone*

Pour tenter d'accroître la robustesse du système, nous avons intégré dans le processus de vérification une technique de synchronisation des alignements des observations acoustiques dans le modèle client et dans le modèle du monde. Selon cette approche [M+99], la séquence d'états dans les deux modèles est exactement la même et est, en l'occurence, définie par l'alignement dans le modèle du monde.

La figure 4 montre les résultats avec adaptation dans le cas de l'utilisation d'un alignement synchrone (sur le modèle Ω) pour l'apprentissage et le décodage.

Il est intéressant de noter que les performances en terme de courbes DET sont similaires à celles observées précédemment (figure 3), mais que la dérive du seuil est considérablement réduite, avec une fluctuation du seuil optimal de l'ordre de 5 % seulement autour du seuil théorique. En outre, le temps nécessaire à une vérification est quasiment divisé par deux, car il suffit d'effectuer un seul décodage Viterbi au lieu de deux. Ces résultats supplémentaires confirment donc tout l'intérêt de la technique d'alignement synchrone pour la vérification du locuteur.



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	-0.044	0	0.035	0.069

FIG. 4 – *Dérive des seuils avec une adaptation Bayesienne du modèle Ω avec $\gamma = 2/3$ et alignement synchrone sur Ω à l'apprentissage et au décodage.*

6. CONCLUSIONS

Nos travaux tendent à mettre en évidence l'apport des techniques d'adaptation à différents niveaux de l'apprentissage des modèles de locuteur. Nos expériences illustrent l'intérêt d'adapter le modèle client à partir d'un modèle indépendant du locuteur. Elles valident également l'utilisation d'un apprentissage incrémental permettant de remettre à jour de façon incrémentale le modèle du client à partir des énoncés prononcés en phase opérationnelle, sans avoir à stocker l'ensemble des données acoustiques correspondantes. Enfin, nous confirmons l'intérêt de l'alignement synchrone qui semble contribuer à faciliter le réglage et le suivi des seuils en apprentissage incrémental.

Une des étapes suivantes consiste à étendre cette étude au cas de l'apprentissage non-supervisé, c'est-à-dire sans savoir a priori si les énoncés d'apprentissage ont effectivement été produits ou non par le client.

REMERCIEMENTS

Ce travail est financé par l'OFES (Office Fédéral de l'Education et de la Science), project n 97.0494-2 et par la CE (Commission Européenne) Telematics Programme LE4 (project 8369).

RÉFÉRENCES

[B+99] F. Bimbot et al. An overview of the picasso project research activities in speaker verification for telephone applications. In *6th european conference on speech communication and technology — eurospeech'99*, volume 5, pages 1963–1966, Budapest, Hungary, September 5–10 1999.

[F+00] C. Fredouille et al. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In *ICASSP2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 5–9 2000.

[GL94] J. L. Gauvain and C.-H. Lee. Maximum a posterionri estimation for multivariate gaussian mixture observation of markov chains. In *IEEE Transactions on Speech Audio Processing*, volume 2, pages 291–298, April 1994.

[M+99] J. Mariéthoz et al. Client / world model synchronous alignement for speaker verification. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, Budapest, Hungary, September 5–10 1999.

[MC99] C. Mokbel and O. Collin. Incremental enrollment of speech recognizers. In *ICASSP'99*, 1999.

[Mok98] C. Mokbel. Incremental enrollment. PICASSO WP5 Delivrable D5.1, December 1998.

[MP97] A. Martin and M. Przybocki. The det curve in assessment of detection task performance. In *Eurospeech 97*, volume 4, pages 1895–1898, 1997.

[Rey97] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech 97*, volume 2, pages 963–966, 1997.

Extension de la recherche de meilleure base pour la décomposition en paquets d'ondelettes. Application à l'analyse en sous-bandes de la parole

Gilles Gonon, Silvio Montresor, Marc Baudry

Laboratoire d'Informatique de l'Université du Maine
Université du Maine, 72085 Le Mans Cédex 9, France
Mèl: Gilles.Gonon / Silvio.Montresor / Marc.Baudry@lium.univ-lemans.fr

Abstract

In the audio signal processing area (coding or restauration), subband analysis shows to be an efficient tool. Extensions of the dyadic basis usually used in Best Basis search have been proposed in former work. This article review these extensions and presents an easy way to construct the filter bank associated with such basis. The filters, designed from any usual Quadrature Mirror Filters and preserving their reconstruction properties, allow to generate the father of two adjacent subbands not coming from the same father in the dyadic decomposition and thus to perform the entropic test between these subbands, which is not otherwise possible. We then apply this new Best Basis on a speech signal wavelet packet decomposition.

1. Introduction

L'algorithme du choix de la meilleure base permet de représenter un signal sur une base d'ondelettes de manière optimale au sens d'une fonction de coût, comme par exemple l'entropie de Shannon [eMW92]. Il est ainsi possible de trouver une partition de l'axe fréquentiel fournissant une analyse pertinente du signal au sens où l'on cherche à en isoler les différentes composantes fréquentielles.

Dans le cas des paquets d'ondelettes, la meilleure base est obtenue à partir d'une décomposition dyadique du signal. Cependant, il subsiste un problème se traduisant par l'introduction de ruptures artificielles générées par la structure de la décomposition [AMB97]. Il est en effet impossible de réunir deux paquets de coefficients correspondant à des bandes de fréquences contigües mais ne provenant pas du même père. La figure 1 montre quelles sont ces ruptures. En suivant les notations de la figure 1, le test entropique ne permet pas de réunir les nœuds $(2,1)$ et $(2,2)$. Cette rupture artificielle est la première due à la structure dyadique. Elle reste présente dans la suite de la décomposition et se répète à chaque niveau, à une échelle inférieure.

En faisant l'analogie de la transformée en paquets d'ondelettes discrète avec une décomposition en sous-bandes par des filtres QMF, il est envisageable de trouver la sous-bande équivalente au père de deux nœuds provenant normalement de deux pères différents dans la décomposition dyadique. Il faut pour cela considérer différemment le problème et ajouter un degré de liberté à l'analyse en sortant de la structure QMF des filtres utilisés, tout en conservant la propriété de reconstruction parfaite.

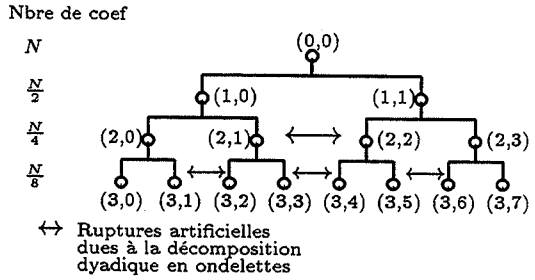


Figure 1: Mise en évidence des ruptures non supprimées par les algorithmes de recherche de meilleure base

Le problème est donc maintenant celui de la construction d'un banc de filtre non-uniforme à reconstruction parfaite.

Le paragraphe 2 présente une méthode pour réaliser le banc de filtres permettant de retrouver le père des deux bandes centrales. Le paragraphe 3 propose une méthode permettant l'analyse du signal en travaillant à échantillonnage critique. Le paragraphe 4 donne un exemple de banc généré qui est alors appliqué à la recherche de meilleure base sur un signal de parole. Enfin le paragraphe 5 propose quelques perspectives de recherche et d'applications de la méthode.

2. Construction du banc de filtres

Le banc de filtres permettant l'élimination de la rupture doit partitionner la bande fréquentielle suivant les intervalles $[0, \frac{\pi}{4}]$, $[\frac{\pi}{4}, \frac{3\pi}{4}]$ et $[\frac{3\pi}{4}, \pi]$ (π correspondant ici à la fréquence de Nyquist). Dans cette partition de l'axe des pulsations, en nous référant à la figure 1, les sous-bandes $[0, \frac{\pi}{4}]$ et $[\frac{3\pi}{4}, \pi]$ doivent correspondre respectivement aux coefficients relatifs aux nœuds $(2,0)$ et $(2,3)$ tandis que la bande $[\frac{\pi}{4}, \frac{3\pi}{4}]$ correspond au père des nœuds $(2,1)$ et $(2,2)$.

Le banc de filtres désiré doit garder les propriétés des filtres QMF, à savoir leurs réponses fréquentielles permettant d'obtenir la reconstruction parfaite du signal. Pour garder ces propriétés, la méthode utilisée consiste à sur-échantillonner les réponses des filtres QMF. Ceci vient du fait que pour passer à la profondeur inférieure, le signal est sous-échantillonné. La démarche utilisée consiste donc à sur-échantillonner les filtres au

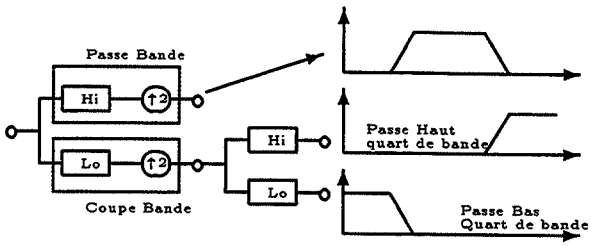


Figure 2: Construction du banc de filtres par filtrage par les QMF

lieu de sous-échantillonner le signal.

Les deux filtres obtenus par sur-échantillonnage vérifient aisément les propriétés de reconstruction parfaite puisque l'on ne fait qu'ajouter des zéros. Le filtre passe-bas devient après sur-échantillonnage un filtre *coupe-bande*. Sa bande passante est $[0;\frac{\pi}{4}] \cup [\frac{3\pi}{4};\pi]$. Le passe-haut devient lui un passe-bande couvrant la bande $[\frac{\pi}{4};\frac{3\pi}{4}]$.

Le banc de filtres ainsi obtenu a bien l'allure souhaitée mais il reste à scinder le coupe-bande en deux filtres équivalents aux filtres passe-bas et passe-haut $\frac{1}{4}$ de bande. À cette fin, deux méthodes sont proposées :

- La première méthode consiste à moduler la version analytique du passe-bande par $\pm\frac{\pi}{2}$. Les filtres obtenus par chacune de ces modulations complexes sont réels et correspondent aux deux filtres souhaités. Cette méthode, bien que cohérente avec la forme QMF du filtre passe-haut ne fournit pas la reconstruction parfaite lorsque l'on travaille à échantillonnage critique car elle ne tient compte que d'un des deux filtres QMF, le passe-haut. Elle présente l'avantage de générer 3 filtres de même longueur.
- La seconde méthode consiste à filtrer le passe-bas sur-échantillonné (coupe-bande) par les deux filtres QMF. La division du filtre ainsi obtenue est garantie à reconstruction parfaite de par la nature des filtres QMF. Cette méthode, illustrée à la figure 2, reste cohérente avec la décomposition dyadique du signal car les seuls filtres utilisés sont les filtres QMF.

Nous conservons dans la suite le banc généré par la deuxième méthode. Par rapport à la décomposition en paquets d'ondelettes, le banc de filtres permettant de passer directement à la profondeur 2 est obtenu de manière très similaire, comme le montre la figure 3. Dans ce cas là, le passe-haut sur-échantillonné est lui-même filtré par les deux QMF, donnant les sous-bandes correspondant aux nœuds $(2,1)$ et $(2,2)$. L'égalité des sous-bandes obtenues par les deux méthodes montre l'équivalence entre le banc de filtres généré et la décomposition en ondelettes. Le filtre passe-bande correspond ainsi au père des nœuds (2,1) et (2,2) de la transformée en ondelettes.

Nous allons maintenant présenter le cas de l'analyse à échantillonnage critique, intéressant pour les applications de codage en sous-bandes et de compression, pour lesquelles il ne doit pas y avoir d'augmentation

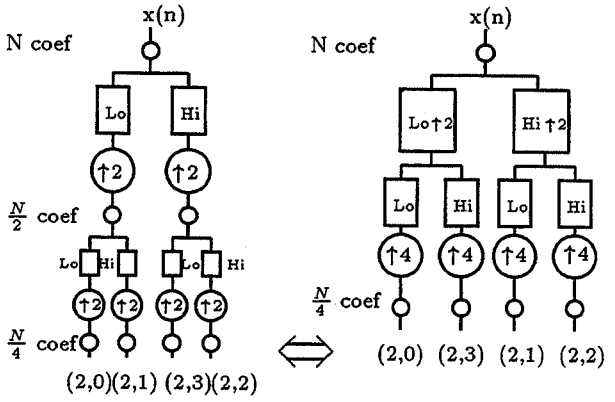


Figure 3: Obtention du banc de filtres équivalent à la profondeur 2 de la décomposition en paquets d'ondelettes.

du nombre de coefficients résultant de l'analyse.

3. Analyse à échantillonnage critique

Cette section présente les problèmes survenant lorsque l'on veut travailler à échantillonnage critique et propose une solution pour le sous-échantillonnage de la sous-bande $[\frac{\pi}{4},\frac{3\pi}{4}]$.

Le banc de filtres généré à la section précédente est un banc non-uniforme et le sous-échantillonnage de la bande du centre n'est pas directement possible car les conditions relatives au changement de fréquence d'échantillonnage (ou de résolution) ne sont pas remplies. Les résultats relatifs aux opérations de sur et sous-échantillonnage sont présentés dans [CR83] et [VK95].

3.1. Nécessité de la modulation

Dans le domaine de la compression, il est important de travailler à échantillonnage critique pour ne pas introduire de redondance sur le signal lors de l'analyse. Ainsi chaque sous-bande va être analysée à sa résolution minimale, et le sous-échantillonage se fait de sorte que le spectre des sous-bandes soit élargi à toute la bande ($[0,\pi]$). Dans le cas des bancs de filtres uniformes, c'est-à-dire composés de filtres de même largeur fréquentielle, le sous-échantillonnage ne pose aucun problème. Prenons le cas général d'un banc de M-filtres partitionnant l'axe fréquentiel en M bandes $[\frac{k\pi}{M},\frac{(k+1)\pi}{M}]$, $k=0,\cdots,M-1$, qui est le cas des décompositions dyadiques ($M=2^D$). Après l'étape de sous-échantillonnage d'un facteur M les bandes paires s'élargissent à tout l'axe fréquentiel, les bandes impaires aussi mais leurs spectres se renversent et il faut donc leur appliquer l'opérateur Mirroir $(-1)^n$.

Dans le cas du banc de filtres proposé ici, la réunion des 2 sous-bandes d'un banc uniforme entraine que ce dernier n'est plus uniforme. Aussi, les deux sous-bandes $[0,\frac{\pi}{4}]$, et $[\frac{3\pi}{4},\pi]$ peuvent être sous-échantillonnées d'un facteur 4 sans problème. Par contre, pour la sous-bande $[\frac{\pi}{4},\frac{3\pi}{4}]$, le sous-échantillonnage n'est pas directement réalisable car dans ce cas là une partie du spectre se replie sur elle-même. Il faut donc au préalable moduler cette

sous-bande par $\frac{\pi}{4}$ pour la ramener dans l'intervalle $[0, \frac{\pi}{2}]$, après quoi elle peut être sous-échantillonnée d'un facteur 2. La figure 4 illustre les problèmes de l'échantillonnage critique liés au recouvrement et au renversement des sous-bandes impaires.

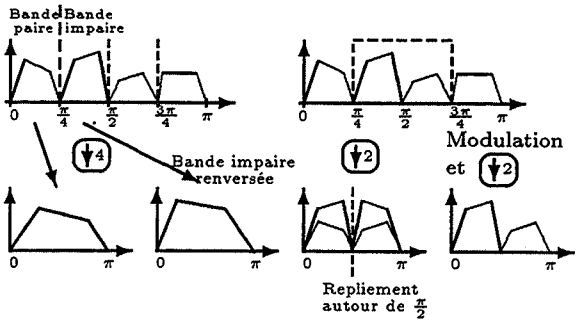


Figure 4: Sous-échantillonnage pour un banc de filtres uniforme et pour la sous-bande du centre du banc de filtres proposé.

3.2. *Modulation de la sous-bande centrale*

La modulation fréquemment utilisée pour le codage est la modulation Single-Side Band, qui fournit des coefficients réels et n'introduit ainsi pas de redondance par passage au domaine complexe. Elle est présentée dans [CR83] et va permettre dans notre cas de sous-échantillonner la bande du centre d'un facteur 2.

La figure 5 illustre les chaînes de modulation et de démodulation. La modulation porte sur le signal filtré par le passe-bande $[\frac{\pi}{4}, \frac{3\pi}{4}]$ et la sous-bande démodulée est synthétisée par le même filtre renversé.

Une fois la bande du centre modulée et sous-échantillonnée, il est possible de poursuivre sur cette sous-bande une décomposition dyadique classique ou de recommencer la décomposition présentée ici.

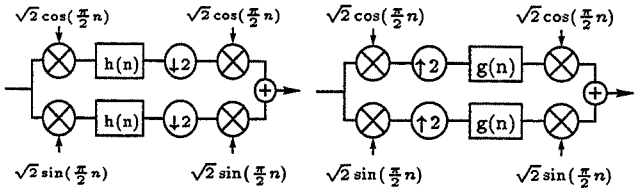


Figure 5: Modulation Single Side Band de la bande du centre et sous-échantillonage critique

4. Simulations

4.1. *Visualisation du banc de filtres*

La figure 6 présente le banc de filtres QMF et le banc pseudo-QMF réalisés à partir des ondelettes "Symmlets". Ces ondelettes permettent la reconstruction parfaite du signal, c'est-à-dire que le recouvrement généré aux étapes de sous et sur-échantillonnage s'annule à la synthèse, car les filtres ont des bandes de transition non nulles. Les Symmlets présentent des propriétés intéressantes de régularités et de symétrie qui en font des ondelettes intéressantes pour la recherche de meilleure base.

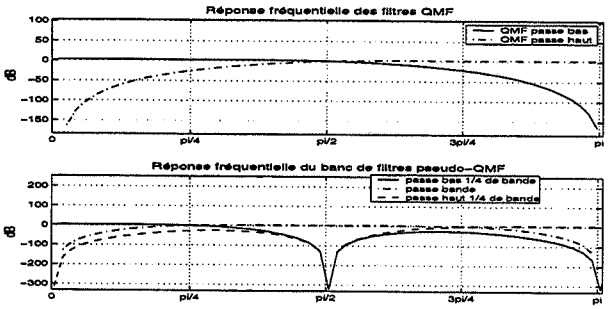


Figure 6: Allure du banc de filtres construit à partir des filtres QMF issus des Symmlets

4.2. *Recherche de la meilleure base appliquée à un signal de parole*

Dans la décomposition en paquets d'ondelettes, la meilleure base est souvent recherchée avec un algorithme de type "bottom-up" consistant à comparer la somme des entropies de 2 nœuds *fils* avec celle de leur *père* et à garder l'étage donnant l'entropie la plus faible, en remontant l'arborescence jusqu'au signal initial(profondeur 0). Pour prendre en compte de la bande du centre, il faut introduire quelques modifications dans le test. Le test porte en effet non plus sur 2 *pères* mais sur 3. Il faut donc une condition supplémentaire permettant de retenir la bande centrale si besoin est. Ainsi la bande du centre sera retenue si son entropie est inférieure à l'entropie de la somme de ses *fils* d'une part, et si elle est inférieure à chacune des entropies des 2 *frères* adjacents issus de la décomposition dydadique d'autre part.

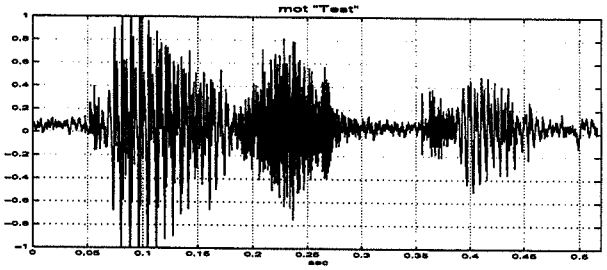


Figure 7: Allure temporelle du signal de simulation

Nous avons alors effectué un test comparatif des deux algorithmes sur le mot "test" à différentes fréquences d'échantillonnage (44.1kHz et 16kHz) et à différentes profondeurs de décomposition (4 et 8). Les figures 8 et 9 illustrent les résultats obtenus à la profondeur 4 pour les paquets d'ondelettes issus de la décomposition dyadique complétés par la bande centrale. La profondeur 8 étant difficilement représentable sur un graphique, le tableau 1 résume les principaux résul-

tats.

Lorsque le signal est échantillonné à 44.1kHz, à la profondeur 4, la meilleure base dyadique obtenue est le dernier étage de la décomposition, tandis que le test sur la bande du centre permet de regrouper les nœuds (4,1) et (4,2), offrant une diminution de l'entropie de quelques pourcents. La figure 8 montre la meilleure base obtenue par les deux méthodes, ainsi que le gain entropique résultant du regroupement des nœuds (4,1) et (4,2). La figure 9 montre les entropies des différents paquets générés.

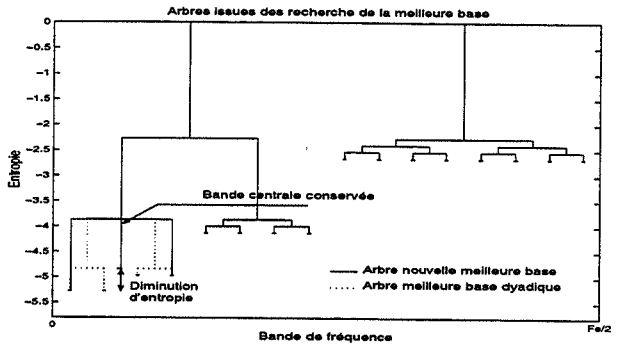


Figure 8: Meilleure base obtenue à la profondeur 4.

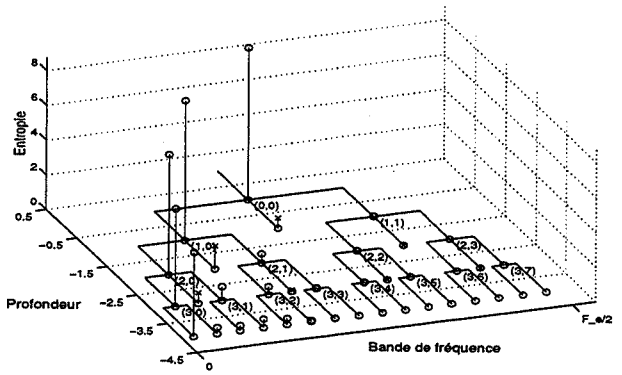


Figure 9: Valeurs des entropies des différents paquets d'ondelettes, et des bandes centrales générées.

Lorsque le signal est échantillonné à 16 kHz, la décomposition à la profondeur 4 choisit les nœuds (3,3) et (3,4) comme terminaux, nœuds qui correspondent à la même zone de fréquence que la bande centrale conservée dans le cas précédent, mais décalée par le sous-échantillonnage.

Pour une décomposition allant jusqu'à la profondeur 8, le tableau 1 montre le nombre de bandes centrales retenues comme nœuds terminaux aux différentes profondeurs après recherche de la meilleure base et le compare au nombre de sous bandes retenues dans la décomposition dyadique. Le nombre de sous-bandes centrales retenues n'est jamais négligeable devant le nombre de sous-bandes retenues par l'algorithme dyadique, ce qui justifie l'ajout du test en vue de la segmentation optimale au sens de la minimisation de l'entropie.

Table 1: Comparaison des nombres de nœuds terminaux conservés (mot "test"; décomposition de profondeur 8).

F_e	Profondeur	Bandes centrales	Bandes dyadiques
$44.1kHz$	7	9	80
	6	5	22
	5	1	7
	4	0	1
$16kHz$	7	14	50
	6	5	15
	5	3	5
	4	1	1
	3	0	1

5. Conclusion et perspectives

Le banc de filtres généré est construit à partir des filtres QMF et propose une nouvelle décomposition du signal permettant de tester des ruptures artificielles dues à la structure dyadique de la décompostion en paquets d'ondelettes. Le test entropique permettant de retenir la sous bande centrale générée vient s'ajouter au test usuel dans un algorithme de type "bottom-up", en retenant la sous-bande centrale si son entropie est inférieure à celles de ces *frères* dyadiques adjacents ainsi qu'à la somme de ces deux *fils*. Les résultats obtenus montrent que l'ajout d'un test dans la recherche de meilleure base permet d'améliorer la segmentation du signal au sens de la minimisation de l'entropie du signal.

Dans de nombreuses applications audios telles que le codage de la parole en bande large, il est important de disposer d'une structure d'analyse pouvant s'adapter au signal. Aussi la méthode proposée permet une meilleure adaptation que les paquets d'ondelettes au sens où elle ajoute un test par rapport à la décomposition dyadique. Nous envisageons donc maintenant d'appliquer cette analyse au codage de la parole en bande élargie ou de la musique.

Bibliographie

[AMB97] Imad Abdallah, Silvio Montrésor, and Marc Baudry. Construction de banc de filtres non uniformes à partir des paquets d'ondelettes. 16^e *Colloque du GRETSI*, 1997.

[CR83] Ronald Crochiere and Lawrence Rabiner. *Multirate Digital Signal Processing.* Prentice-Hall Signal processing series, 1983.

[eMW92] R.R. Coifman et M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transaction on Information Theory, Vol. 38(2), pp. 713-718, Mars 1992*, 38(2), March 1992.

[VK95] Martin Vetterly and Jelena Kovačević. *Wavelets and subbands coding.* Prentice-Hall, 1995.

Robustesse de la vérification du locuteur par mot de passe personnalisé

Bruno JACOB [△], *Johnny MARIETHOZ* [▼], *Guillaume GRAVIER* [▽], *Frédéric BIMBOT* [△]

[△] IRISA (CNRS/INRIA), Campus Universitaire de Beaulieu, 35042 Rennes Cedex, http://www.irisa.fr
[▼] IDIAP, rue du Simplon 4, Case postale 592, CH-1920 Martigny, http://www.idiap.ch
[▽] ENST, 46 rue Barrault - Paris 75634 Cedex 13, http://www.enst.fr

E-mail: {bimbot,bjacob}@irisa.fr - marietho@idiap.ch - gravier@tsi.enst.fr

Abstract

This paper presents a speaker verification approach using customized password (i.e passwords chosen by the client). In this context, the issue consists in estimating a speaker-independent (world) model for the password, using speech utterances from a single speaker. For this purpose, we use a universal (speaker-independent) speech model from which we derive a model of the password on the basis of the decoded (graph of) units. Our results tend to show that the quality of the transcription influences strongly the performance of the system.

1. Introduction

Le travail présenté dans cet article porte sur la vérification du locuteur au téléphone dans le contexte de la sécurisation de transactions ou d'accès à des services à caractère commercial. Il s'inscrit dans le cadre du projet européen Telematics PICASSO [1] [BBB+99].

Pour ce type d'applications, la technologie couramment utilisée repose sur des approches *dépendantes du texte*, opérant la vérification du locuteur sur une séquence de mots *fixe* et invariable (par exemple, une séquence de chiffres correspondant au numéro de compte ou d'abonné du client). Pour réduire les efforts de mémoire de la part de l'utilisateur, ainsi que les risques de fraude, certains profils d'applications (dits à *texte prompté*) font appel à des séquences aléatoires de mots issus d'un petit vocabulaire, commun à tous les clients, que le locuteur doit prononcer en réponse à un *prompt* [Dod85]. Cependant, ces approches sont toutes deux vulnérables à l'utilisation de parole pré-enregistrée, dès lors que l'imposteur est capable de procéder à de la synthèse par concaténation, même si la qualité de la concaténation n'est pas très bonne (voir [LB99] dans le cas de concaténation de mots).

L'approche par *mot de passe personnalisé* (MPP) a pour objectif de remédier, au moins partiellement, aux difficultés qui viennent d'être exposées. Pour ce profil d'application, le client choisit lui même l'énoncé sur lequel s'effectue la vérification. Le premier avantage de cette approche réside en ce qu'elle introduit un niveau supplémentaire de protection car les imposteurs intentionnels potentiels doivent d'abord avoir connaissance du mot de passe avant de pouvoir tenter une imposture (ce qui va de pair avec une impression de sécurité accrue de la part des utilisateurs). Par ailleurs, cette approche est plus ergonomique car elle réduit les difficultés de mémorisation pour l'utilisateur.

Nous présentons dans cet article une série de premiers résultats d'expériences simulant un procédé de vérification du locuteur à partir de mots de passe personnalisés, produits dans le cadre du projet PICASSO. Nous exposons tout d'abord le cadre théorique du problème de la vérification du locuteur. Puis nous introduisons la problématique essentielle de la vérification par mot de passe personnalisé, à savoir l'inférence d'un modèle acoustique *indépendant* du locuteur à partir de l'énoncé d'un *seul* locuteur. Nous décrivons une technique possible pour résoudre cette question, qui utilise un modèle de parole universel (indépendant du texte et du locuteur). Nous présentons finalement une série de résultats expérimentaux qui permettent notamment de comparer l'approche à base de mot de passe personnalisé à une approche dépendante du texte sur les mêmes énoncés.

2. Cadre théorique

2.1. Modèle probabiliste

L'approche utilisée dans l'ensemble de cet article s'appuie sur un formalisme probabiliste du problème de la vérification. Pour un énoncé de test noté Y prononcé par un locuteur proclamant l'identité X, on calcule le logarithme du rapport de vraisemblance :

$$S_X(Y) = log\left(\frac{\hat{\mathcal{P}}(Y|X)}{\hat{\mathcal{P}}(Y|\bar{X})}\right)$$

où $\hat{\mathcal{P}}(Y|X)$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par le locuteur proclamé et où $\hat{\mathcal{P}}(Y|\bar{X})$ représente la vraisemblance de l'énoncé sous l'hypothèse qu'il a été prononcé par un autre locuteur.

Le modèle probabiliste correspondant à X (dit *Modèle du Client*) est estimé à partir de données d'apprentissage composées d'énoncés prononcés par X. Le modèle correspondant à $\bar{X}$ (dit *Modèle Non-Client*) est obtenu à partir d'énoncés semblables prononcés par d'autres locuteurs. Habituellement, le modèle du non-client est le même pour tout les clients, et on le désigne couramment par *Modèle du Monde* (noté Ω). Dans le cas du mot de passe personnalisé, le modèle Ω dépend du mot de passe du locuteur et donc du locuteur. Il faut donc disposer d'un *modèle Client* et

1. partenaires : les opérateurs KPN-Telecom (*NL*), Fortis (*NL*), Swisscom (*CH*), les laboratoires de recherche ENST (*F*), IDIAP (*CH*), KPN-Research (*NL*), KTH (*S*), KUN (*N*), les banques de l'UBS-Ubilab (*CH*) et la société de technologie vocale Vocalis (*UK*).

d'un *modèle du Monde* pour chaque locuteur.

Dans les travaux décrits ici, les modèles probabilistes utilisés sont des MMC (Modèles de Markov Cachés) dont les fonctions d'émission des états sont des mélanges de distributions Gaussiennes.

2.2. *Décision, types d'erreur et mesure des performances*

Dans les applications où il s'agit de prendre une décision binaire d'acceptation ou de rejet de l'identité proclamée, le score S_x est comparé à un seuil de décision choisi de façon à optimiser les performances du système dans une condition de fonctionnement particulière. Cette condition de fonctionnement est spécifiée par le rapport des coûts associés aux deux types d'erreur possibles : *faux rejet*, si un client authentique est rejeté par le système et *fausse acceptation* si un imposteur n'est pas détecté.

Les performances des approches décrites dans cet article sont présentées essentiellement sous forme de courbes DET [MP97] qui indiquent les caractéristiques du système en terme de pouvoir de séparation des clients et des imposteurs : plus la courbe DET est proche de l'origine, meilleure est la séparation apportée par le système. Le point situé à l'intersection de la courbe DET et de la première bissectrice correspond à l'EER (Equal Error Rate) c'est-à-dire le taux d'erreur du système quand les fausses acceptations sont égales aux faux rejets.

3. Problématiques spécifiques au mot de passe personnalisé

La mise en oeuvre du procédé de vérification nécessite d'être en mesure d'estimer le Modèle du Client (X) et le Modèle du Monde (Ω) du mot de passe. Le modèle X caractérise la façon dont le client prononce son mot de passe, ce qui peut être déduit des énoncés d'apprentissage produits pendant la phase d'apprentissage. En revanche, le modèle Ω doit caractériser la façon dont l'ensemble des locuteurs sont susceptibles de prononcer ce même mot alors que l'on ne dispose pas de tels exemples en pratique. Ceci pose donc le problème d'être capable d'inférer un modèle acoustique indépendant du locuteur à partir d'un énoncé produit par un seul locuteur.

Par ailleurs, même si le Modèle du Client peut, pour sa part, être estimé à partir des données d'apprentissage, le volume de données disponibles pour l'estimation est, en pratique, très insuffisant par rapport au volume de données nécessaires pour un apprentissage par maximum de vraisemblance complet.

Le paragraphe suivant expose les réponses que nous apportons à ces deux difficultés.

4. Approche

L'approche adoptée repose sur l'utilisation d'un Modèle de Parole Universel (noté U) représentant les propriétés acoustiques de la langue considérée, indépendamment du locuteur et du texte.

4.1. *Modèle de Parole Universel*

Le Modèle de Parole Universel U est un réseau ergodique dont les états sont des MMCs gauche-droite, que nous dénoterons par MMC_i. On modélise ainsi un réseau d'unités acoustiques (notées U_i) dont les transitions successives sont régies par un modèle bigramme. Ces unités U_i sont indépendantes du locuteur.

Le modèle U est utilisé pour transcrire en unités de parole U_i l'énoncé d'apprentissage (noté Y_0) du mot de passe produit par le client. Ce modèle U fournit l'ensemble des N-meilleures solutions consistant en N séquences d'unités U_i. Cet ensemble de solutions est ensuite factorisé sous forme d'un graphe minimal gauche-droite : $\mathcal{G}(Y_0)$. Ce graphe peut être compris comme un ensemble d'hypothèses de décodage de l'énoncé Y_0 dans le système de représentation linguistique à base des unités U_i.

4.2. *Modèle du Monde*

Le Modèle du Monde est construit à partir du graphe $\mathcal{G}(Y_0)$ en y compilant les MMC_i acoustiques indépendants du locuteur correspondant aux unités U_i dans le graphe. Les états des MMC_i sont représentés par des densités de probabilité multigaussiennes. Le modèle Ω représente donc un graphe d'hypothèses de modèles acoustiques pour le mot Y_0 en mode indépendant du locuteur.

4.3. *Modèle du Client*

La création du modèle du Client s'effectue en deux phases :

1. Dans une première étape, le Modèle du Client est construit à partir du graphe $\mathcal{G}(Y_0)$ en y compilant les MMC_i^{mono} acoustiques indépendants du locuteur correspondant aux unités U_i dans le graphe. Les états des MMC_i^{mono} sont représentés par des densités de probabilité mono-gaussiennes. On peut voir les MMC_i^{mono} comme des versions simplifiées des MMC_i. Le modèle X_{init} ainsi obtenu sert de modèle d'initialisation au modèle X du Client.
2. On entraîne ensuite (seulement) les moyennes des états du modèle initial X_{init} au sens du maximum de vraisemblance à partir d'une ou de plusieurs répétitions du mot de passe par le client. Les variances du modèle Client restent égales à celles du modèle X_{init}. Une prochaine étape de notre travail consistera à effectuer cet apprentissage par des techniques Bayésiennes [MM99].

La figure 1 illustre les étapes de ce processus.

5. Protocole expérimental et expériences

5.1. *Base de données*

La base de données sur laquelle nous avons réalisé nos expériences est un sous-ensemble de la base de données Polyvar de l'IDIAP. Elle est composée d'énoncés correspondant à 17 mots de commande d'une application touristique de la ville de Martigny en Suisse : annulation, cinéma, Corso, galerie du Manoir, guide, manifestation, mode d'emploi, précédent, suivant, casino, concert, exposition, Giannada, Louis Moret, message, musée, quitter. Dans cette étude, nous avons sélectionné une population de 19 locuteurs (12 Hommes et 7 Femmes).

L'ensemble d'apprentissage contient les 5 premières

répétitions de chaque mot de passe de chaque locuteur.

L'ensemble de test est composé d'environ 18000 énoncés dont 1/3 sont des accès authentiques et 2/3 sont des accès d'imposteurs. Les accès imposteurs sont les mêmes énoncés que les accès clients, mais testés contre d'autres identités.

5.2. *Protocole d'évaluation*

Dans nos expériences, nous avons considéré qu'un mot de commande était un mot de passe hypothétique. Il faut remarquer que ces mots de passe ne sont formés que d'un très petit nombre de syllabes : la décision d'accepter ou non un locuteur comme client doit donc s'effectuer sur un temps très court et rend la tâche de vérification singulièrement difficile.

Les modèles acoustiques des unités de parole U_i représentent des phonèmes hors contexte appris avec la base PolyPhone Suisse Romand. Chaque phonème est composé de trois états émetteurs. Les lois d'émission sont formées d'un mélange de trois gaussiennes pour les MMC_i, et d'une mono-gaussienne pour les MMC_i^{mono}.

Le graphe $\mathcal{G}(Y_0)$ est obtenu à partir des 3 meilleures solutions pour l'énoncé Y_0.

Pour évaluer cette approche, nous l'avons comparée à une expérience de référence utilisant explicitement des exemples acoustiques de chaque mot de passe. L'approche de référence s'apparente donc à une méthode dépendante du texte. Dans cette configuration, on dispose de données acoustiques multi-locuteurs et on est en mesure d'estimer le modèle du *Monde* à partir de celles-ci. En pratique, les énoncés utilisés pour estimer le modèle du Monde proviennent d'un ensemble de locuteurs distincts des 19 locuteurs utilisés dans nos tests.

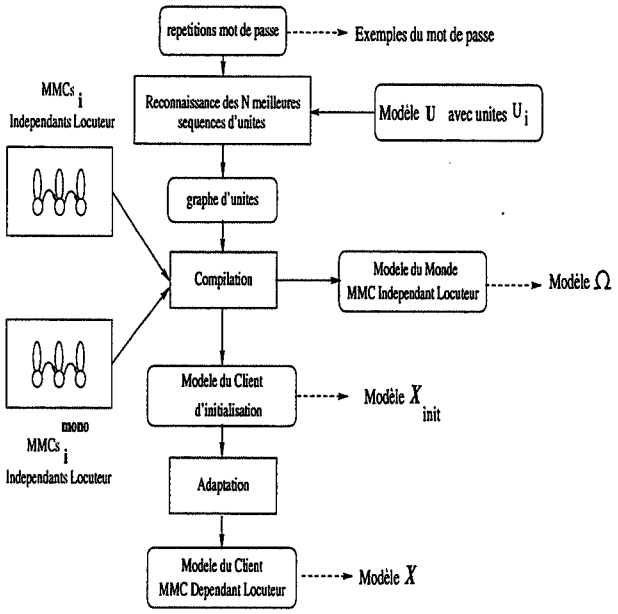


FIG. 1 – *Synopsis de la création du mot de passe personnalisé*

5.3. *Résultats*

Comme on l'observe sur les courbes DET de la figure 2 et dans la table 1, les performances de l'approche par mot de passe personnalisé (MPP) s'améliorent avec le nombre de répétitions d'apprentissage du mot de passe, mais elles plafonnent au-delà de 4 répétitions.

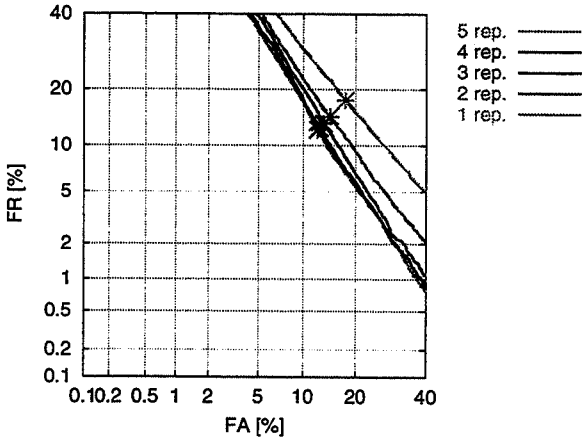


FIG. 2 – *Résultats de la méthode MPP en fonction du nombre de répétitions.*

TAB. 1 – *EER de la méthode MPP en fonction du nombre de répétitions.*

Nb répétitions	1	2	3	4	5
EER %	17,7	14,5	13,4	12,6	12,3

La figure 3 permet de comparer l'approche à base de mot de passe personnalisé (3 répétitions) et l'approche dépendante du texte (de référence) sur les mêmes mots (3 répétitions également). On constate assez naturellement une performance moindre de la méthode MPP par rapport à la méthode de référence avec un EER de 13,4% pour la méthode testée contre 9,6% pour la méthode de référence. Il semble que le facteur

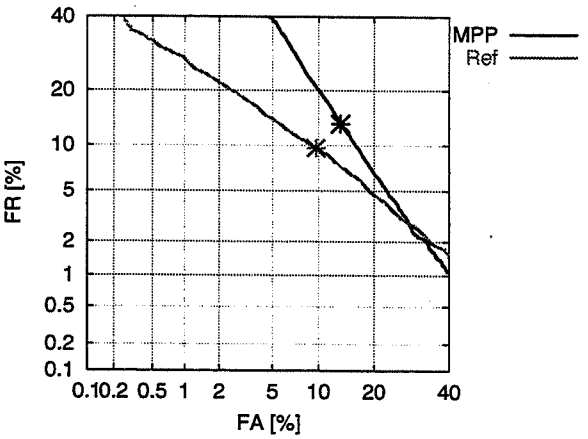


FIG. 3 – *Comparaison de la méthode de référence et de la méthode MPP.*

limitant des deux méthodes provienne beaucoup de la

brièveté des mots de passe utilisés.

Au moins deux raisons peuvent expliquer la différence de performances observées pour les deux méthodes :

1. La plus ou moins bonne adéquation de la *topologie* du modèle Ω obtenu par transcription automatique. En d'autres termes, la pertinence des unités U_i.
2. La meilleure qualité des *lois d'émission* du modèle Ω lorsqu'elles sont estimées à partir de données réelles (dans l'expérience de référence).

C'est pourquoi, nous avons réalisé une expérience complémentaire, où l'étape de transcription automatique dans la méthode MPP est remplacée par une transcription *manuelle* du mot de passe (méthode MPP-TM).

La figure 4 montre que l'utilisation d'une transcription manuelle apporte des amélorations très notables à la méthode par mot de passe personnalisé, conduisant à des performances équivalentes (voire meilleures) qu'avec la méthode de référence (EER de 9,2% contre 9,6% respectivement).

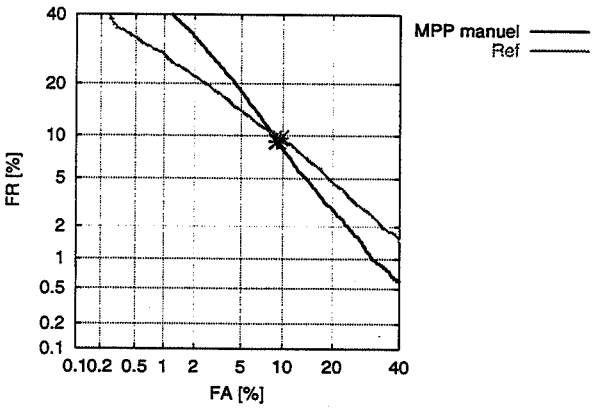


FIG. 4 – *Comparaison de la méthode de référence et de la méthode MPP-TM.*

A l'inverse, nous avons évalué les performances d'une méthode pour laquelle la transcription en unités est effectuée automatiquement mais pour laquelle on procède à la ré-estimation des lois d'émission du modèle Ω à partir de données acoustiques multi-locuteurs (méthode MPP-AC). Cette configuration, sans intérêt pratique, permet néanmoins de mesurer l'importance des données acoustiques.

La figure 5 compare l'approche MPP-AC avec la méthode de référence. Les résultats montrent que la topologie de Ω inférée par décodage des N meilleures solutions n'est pas rédhibitoire, car on peut obtenir des performances comparables à l'expérience de référence si le réapprentissage du Modèle du Monde s'effectue sur des données réelles (EER=10% contre 9,6% respectivement). Cependant, les performances observées sont sensiblement moins bonnes que lorsque l'on dispose de la transcription, même en l'absence de données acoustiques (méthode MPP-TM).

6. CONCLUSIONS

Dans une application de vérification du locuteur par mot de passe personnalisé, l'absence de données acoustique pour estimer le Modèle du Non-Client (du Monde) associé au mot de passe du client, constitue un handicap sérieux. Nos travaux indiquent l'importance cruciale de la qualité du décodage en unités (ici phonétiques) pour envisager d'atteindre des performances comparables á celles des approches dépendantes du texte. Parmi les perspectives d'améliorations, il convient de tester l'utilisation de phonèmes contextuels ou d'unités de parole plus longues que le phonème.

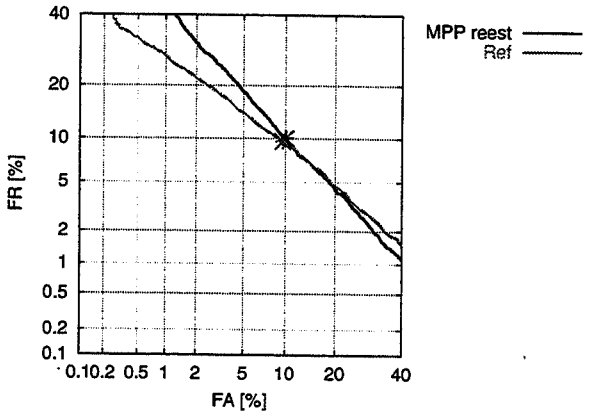


FIG. 5 – *Comparaison de la méthode de référence avec la méthode MPP-AC.*

RÉFÉRENCES

[BBB+99] F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariéthoz, C. Mokbel, and H. Mokbel. An Overview of the PICASSO Project Research Activities in Speaker Verification for Telephone Applications. In *Eurospeech*, volume 5, pages 1963–1966, 1999.

[Dod85] G.R. Doddington. Speaker Recognition – identifying people by their voices. In *Proc. IEEE*, volume 73, pages 1651–1664, March 1985.

[LB99] J. Lindberg and M. Blomberg. Vulnerability in Speaker Verification - A study of Technical Impostor Techniques. In *Eurospeech*, volume 3, pages 1211–1214, 1999.

[MM99] J. Mariéthoz and C. Mokbel. Polyvar Protocol. In *Rapport Interne IDIAP*, 1999.

[MP97] A. Martin and M. Przybocki. The DET curve in assesment of detection task performance. In *Eurospeech*, volume 4, pages 1895–1898, 1997.

Utilisation de Mots de Passe Personnalisés pour la vérification du locuteur

Jamal Kharroubi, Gérard Chollet

ENST-TSI, CNRS URA-820
46, rue Barrault 75634 –PARIS CEDEX 13, France
Tél : ++33 (0)145 81 75 62 - Fax : ++33 (0)145 88 79 35
Mél : kharroub,chollet@tsi.enst.fr - http://www.tsi.enst.fr

ABSTRACT

The level of security of a text-dependent speaker verification system is quite low when a deliberate impostor happens to know the password of the client. Therefore, one solution to increase the security level is to let the client choose his own 'customised' password in the same way he does for his computer account. Therefore, the client may feel better secured.
In this article, the usability of customised password methods in a text-dependent speaker verification system is experimented. Three essential problems are studied:
- How to recognise the client password ?
- Which model should be used to characterise the pronunciation of his password by a registered client ?
- What is the best strategy to accept a client and reject impostors ?
In this article, some solutions are proposed to tackle these problems. The results are quite encouraging.

1. INTRODUCTION

Ce travail s'inscrit dans le cadre du projet Européen PICASSO «Pioneering Caller Authentification for Secure Service Operation » qui répond au besoin de sécurisation dans le domaine des télécommunications et des institutions financières ou de toute autre entreprise exploitant le télé-commerce [Bim99].

La plupart des systèmes présentant un tel service utilisent des codes appelés PIN code (Personal Identification Number) qui permettent d'identifier les clients. Ces codes déterminés par le système et imposés aux clients de l'application, peuvent être une suite de chiffres ou de mots de commande choisis à partir d'un vocabulaire limité. Ils sont en général relativement longs ce qui rend le système difficilement utilisable et peu ergonomique. De plus le niveau de sécurité dans les systèmes de vérification du locuteur devient très faible lorsqu'un imposteur réussit à connaître le code d'un client. Une des solutions pour résoudre ce problème est de permettre au client de choisir son propre mot de passe. C'est alors au système de s'adapter au client en prenant en considération le mot de passe choisi. En conséquence le client sera responsable de la gestion de son compte. C'est la technique des Mots de Passe Personnalisés (MPP).

Un système de vérification du locuteur est une procédure qui permet à partir d'un signal de parole x et d'une identité proclamée λ de décider si le signal x correspond au locuteur λ ou à un imposteur. La majorité des systèmes de vérification du locuteur actuels sont basés sur le principe de test binaire d'hypothèses. Ce principe s'exprime de la façon suivante.

$$\frac{P(x/\lambda)}{P(x/\bar{\lambda})} \underset{H_1}{\overset{H_0}{\gtrless}} \beta \qquad (1)$$

où $\bar{\lambda}$ représente le modèle indépendant du locuteur appelé aussi modèle du monde. H_0 correspond à l'hypothèse que x provient de l'identité proclamée λ, H_1 correspond à l'hypothèse que x ne provient pas de l'identité proclamée λ et β représente le seuil de décision [Ros92]. Ce seuil peut être dépendant ou indépendant du locuteur.

En se basant sur l'approche statistique classique décrite ci-dessus, la mise en place de la technique des mots de passe personnalisés pose 3 problèmes principaux qui correspondent à 3 étapes essentielles :

- le premier problème consiste à reconnaître le mot de passe de chaque client. La difficulté de cette tâche vient du fait que nous ne disposons d'aucune information sur ces mots qui puisse faciliter cette reconnaissance. La solution d'utiliser des mots comme unités acoustiques est écartée parce que les mots de passe appartiennent à un vocabulaire illimité. Pour répondre à ce problème nous proposons d'utiliser des unités segmentales comme les phones, les diphones, etc...

- le deuxième problème est de créer le modèle indépendant du locuteur pour chaque mot de passe. La solution à ce problème dépend principalement du premier dans la mesure où il n'est pas raisonnable d'utiliser une modélisation des unités acoustiques différentes de celle utilisée en reconnaissance. La difficulté de ce problème est de trouver suffisamment de données pour pouvoir entraîner ces modèles.

- le troisième problème consiste à déterminer le modèle du client qui dépend de son mot de passe avec très peu

de données. Dans une application réelle, on ne peut pas faire répéter au client son mot passe de nombreuse fois. La solution que nous proposons est d'adapter les paramètres du modèle du monde obtenu dans la deuxième étape pour construire le modèle du client.

Pour évaluer notre technique, nous avons comparé notre système avec un système de référence où nous supposons que la transcription phonétique du mot de passe d'un client est connue.

Cet article est organisé comme suit : dans la section 2, nous décrivons l'approche que nous proposons. La section 3 présente une description du protocole expérimental et la base de données utilisée dans nos expériences ainsi que les résultats obtenus. Enfin nous concluons et offrons quelques perspectives.

2. APPROCHE PROPOSÉE

Notre système est basé sur trois phases principales : la reconnaissance du mot de passe du client, la détermination du modèle du monde correspondant au mot reconnu et la détermination du modèle du client par adaptation des paramètres du modèle indépendant du locuteur.

Les phones sont des unités acoustiques parmi les plus utilisées en reconnaissance de la parole. C'est pourquoi, dans notre application, nous avons décidé d'utiliser des modèles de phones de la langue française pour la reconnaissance du mot de passe du client. Ces modèles sont des modèles de Markov cachés (HMM) gauche-droite à 3 états et 3 gaussiennes. Pour la détermination du modèle du monde et du modèle du client, les HMM à 3 états et 3 gaussiennes sont remplacés par des HMM à 3 états et 1 gaussienne en raison du peu de données d'apprentissage dont nous disposons pour estimer les paramètres du modèle du client. En général les données d'apprentissage ne dépassent pas 10 secondes de parole par client.

2.1 Apprentissage

Cette phase est composée de la reconnaissance du mot de passe du client et de la modélisation du client.

Une précédente étude menée dans le cadre du projet PICASSO [Mar99] a montré que 5 répétitions du mot de passe sont suffisantes pour modéliser le client lors d'un apprentissage incrémental. Nous modélisons de même le client en utilisant 5 répétitions de son mot de passe.

Reconnaissance du mot de passe du client Chaque répétition du mot de passe est transcrite phonétiquement en utilisant les modèles de phones indépendants du locuteur basés sur des HMM à 3 états et 3 gaussiennes. A chaque répétition nous obtenons une suite de phones, ce qui nous permet d'avoir pour chaque client 5 transcriptions phonétiques représentant son mot de passe.

Notre modèle du monde d'un mot de passe est l'ensemble des modèles de phones à 3 états et 1 gaussienne apparaissant dans chacune des 5 transcriptions du mot de passe.

Modélisation du client Après la reconnaissance des 5 transcriptions possibles d'un mot de passe d'un client, nous procédons à l'estimation des paramètres du modèle du client par adaptation des paramètres du modèle du monde. Seules les moyennes de chacun des modèles de phonèmes constituant le modèle du monde du mot de passe qui sont adaptées en utilisant toutes les occurrences de ce phonème dans les 5 répétitions. Cette adaptation est réalisée suivant un critère de maximum a posteriori (MAP), la moyenne adaptée correspondant à l'état j étant donnée par

$$\hat{\mu}_j = \frac{N_j}{N_j + \tau}\bar{\mu}_j + \frac{\tau}{N_j + \tau}\mu_j \qquad (2)$$

où N_j est le taux d'occupation de l'état j pour les données d'adaptation, soit

$$N_j = \sum_{r=1}^{R}\sum_{t=1}^{T_r} P_j^r(t)$$

La variable P_j^r correspond à la probabilité d'occupation de l'état j à l'instant t pour l'occurrence r du phone considéré. Dans l'équation (2), $\bar{\mu}_j$ est la moyenne globale pour les données d'adaptation tandis que μ_j est la moyenne pour le modèle indépendant du locuteur [Gau94]. Nous prendrons τ égal à 15 dans toutes les expériences présentées dans cet article.

A cette étape de notre approche, nous avons pour chaque client les 5 transcriptions phonétiques de son mot de passe notées T_i (i=1 ... 5) et un modèle du client adapté à partir du modèle du monde correspondant à son mot de passe.

En ce qui concerne le système de référence, pour lequel nous disposons de la transcription phonétique du mot de passe de chaque client. La modélisation du client est réalisée par adaptation des paramètres du modèle du monde en utilisant au moins 5 occurrences de chacun des phonèmes de la transcription phonétique du mot de passe.

2.2 Test

Etant donné une identité proclamée λ qui doit correspondre à un client de l'application et un signal x, nous procédons d'abord à une reconnaissance forcée du signal x avec les 5 transcriptions phonétiques correspondant au mot de passe du client λ en utilisant le modèle du monde correspondant à ce mot de passe. Cette procédure nous permet d'avoir la transcription la plus probable T_{i*} ainsi que la vraisemblance avec le modèle du monde $P(x / \bar{\lambda}, T_{i*})$.

La vraisemblance avec le client $P(x / \lambda, T_{i*})$ est calculée en réalisant une reconnaissance forcée de x avec la transcription du mot de passe T_{i*} en utilisant le modèle du client.

Le score obtenu par le rapport des deux vraisemblances $P(x / \lambda, T_{i*})$ et $P(x / \overline{\lambda}, T_{i*})$ sera comparé à un seuil pour décider si le signal x provient du client λ ou d'un imposteur.

Pour le système de référence, le test suit exactement les mêmes démarches que nous avons cité précédemment en utilisant la vraie transcription du mot de passe du client au lieu des 5 transcriptions possible.

3. EVALUATION EXPERIMENTALE

3.1 Protocole

Dans nos expériences nous avons utilisé un sous-ensemble de la base POLYVAR [Cho96]. Cette base contient 143 locuteurs dont 58 femmes et 85 hommes qui ont enregistré entre 1 et 229 sessions pour un total de 3600. Une session est un enregistrement de 17 mots qui seront considérées comme des mots de passe. La base de données utilisée dans nos expériences est divisée en deux ensembles : développement et évaluation. L'ensemble de développement sert pour le calcule du seuil que nous appliquerons sur l'ensemble d'évaluation. Chaque ensemble contient 19 locuteurs dont 12 hommes et 7 femmes représentant les éventuels clients et chaque client est considéré comme imposteur pour les autres. Les clients ont été choisis parmi les locuteurs qui ont fait plus de 25 sessions.

Pour chaque client, seules les 5 premières répétitions des mots de passe sont utilisées pour l'apprentissage de son modèle.

Dans la phase de test nous avons réalisé ~6500 accès clients et ~11500 accès imposteurs dans les deux ensemble. Les seuils ont été déterminés expérimentalement sur le corpus de développement et appliqués sur le corpus d'évaluation.

Les modèles de phonèmes indépendants du locuteur ont été entraînés sur un sous-ensemble de la .base de données POLYPHONE de 1000 locuteurs dont 500 femmes et 500 hommes [Cho96]. Chaque locuteur a enregistré à travers des lignes téléphoniques 10 phrases phonétiquement équilibrées pour un total de 10000 phrases. Le nombre de phones utilisé est de 34 plus un modèle de silence.

3.1 Résultats

La figure 1 présente deux courbes DET [Mar97] correspondant aux performances du système utilisant la technique MPP et du système de référence sur le corpus de développement. En comparant les deux courbes, on constate que le système de référence a une meilleure performance que le système utilisant la technique MPP. 13% de TEE (Taux d'Egale Erreurs) pour le système de référence contre 15.5% pour le système MPP. Cette différence significative entre les performances des deux systèmes est due principalement à la connaissance à priori de la vraie transcription des mots de passe des clients dans le système de référence qui permet d'avoir plus de données pour modéliser les clients. Cela n'empêche que les résultats obtenus par la technique MPP restent très encourageants. Notons cependant que ces résultats restent inférieurs à ceux obtenus en utilisant des modèles de mots globaux où le TEE est de 4% [Mar99].

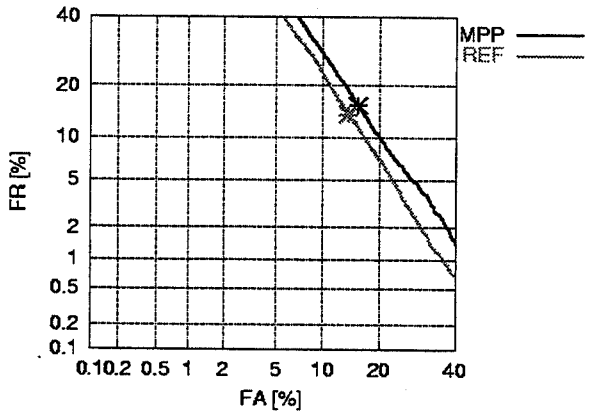


Figure1: Comparaison de la technique des mots de passe personnalisés et le système de référence sur le corpus de développement.

La figure 2 présente les résultats obtenus par le système utilisant la technique MPP et le système de référence sur le corpus d'évaluation. On note toujours que le système de référence a de meilleures performances que le système MPP. En appliquant le seuil correspondant au PF (Point de Fonctionnement) à TEE du corpus de développement sur le corpus d'évaluation, le PF obtenu est très proche du PF à TEE du corpus d'évaluation pour les deux systèmes. Ce qui veut dire que les deux systèmes sont stables.

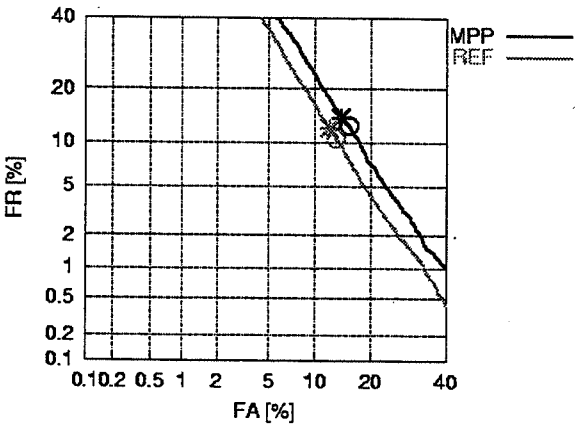


Figure2: Comparaison de la technique des mots de passe personnalisés et le système de référence sur le corpus d'évaluation.

Dans la figure 3, nous présentons les résultats obtenus sur 3 mots de passe de longueur différente choisis parmi les 17 utilisées. Ces mots de passe sont « galerie du manoir, précédent, guide ». Les résultats montrent clairement que plus le mot de passe est long plus les résultats sont meilleurs. On note également que plus le mot de passe est long plus le PF obtenu en appliquant le seuil correspondant au PF à TEE du corpus de développement est proche du PF à TEE du corpus d'évaluation. Ce qui veut dire que plus le mot de passe est long plus le système est stable. Ces résultats peuvent être expliqués par le fait que plus le mot de passe est long plus on a des données qui permettent de bien modéliser les clients.

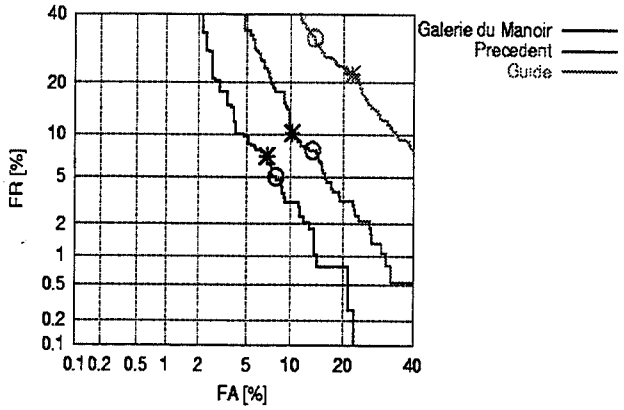


Figure3: Influence de la longueur des mots de passe personnalisés dans le système utilisant la technique MPP.

4. *CONCLUSION*

Les performances d'un système de vérification du locuteur basé sur la technique des mots de passe personnalisés dépend principalement de la qualité du décodeur de reconnaissance de parole utilisé. Un bon choix des unités acoustiques pour la reconnaissance des mots de passe ne peut qu'améliorer les performances du système. Ce travail présente une première étude de la faisabilité de la technique des mots de passe personnalisés. Les résultats obtenus sont très encourageants. Beaucoup de travail reste à faire pour avoir de meilleurs résultats, en particulier en ce qui concerne l'amélioration du décodage acoustico-phonétique ainsi que l'adaptation. On étudiera par la suite les performances en fonction du taux d'adaptation, ainsi que d'autres méthodes d'adaptation comme la régression MLLR. La mise en œuvre de schéma de normalisation, comme la znorm, peut améliorer les performances du système [Gra00]. Cependant, cette technique nécessite de réaliser des accès à l'aide de pseudo-imposteurs ayant prononcé le mot de passe du client. Ceci n'est en pratique pas réalisable et des solutions basées sur la synthèse d'imposteurs devront être envisagées.

REMERCIEMENT

Nous souhaitons remercier Guillaume Gravier pour sa collaboration dans la réalisation de ce travail et lors de la rédaction de cet article. Ce travail a été financé par la Communauté Européenne dans le cadre du projet Telematics PICASSO.

BIBLIOGRAPHIE

[Bim99] Bimbot F. *et al.* (1999), "An overview of the PICASSO project research activities in speaker verification for telephone application", Eurospeech, vol 5, pp 1963-1966.

[Bim97] Bimbot F. *et al.* (1997), "Speaker verification in the telephone network: research activities in the CAVE project, Eurospeech", pp. 971-974.

[Cho96] Chollet G. *et al.* (1996), "Swiss French Polyphone and Polyvar: telephone speech databases to model inter- and intra-speaker variability", Rapport de Recherche de l'IDIAP, RR-96-01.

[Gau94] Gauvain J-l. et Lee C. (1994), "Maximum a posteriori estimation for multivariate Gaussian mixture of Markov chains ", IEEE Trans. on Speech and Audio Processing, vol. 2, n° 2, pp 291-298.

[Gra00] Gravier G., Kharroubi J., Chollet G. (2000), " On the use of prior knowledge in normalization schemes for speaker verification", in Digital Signal Processing: A Review Journal, Academic Press, 2000.

[Mar99] Mariéthoz J., Mokbel C. (1999), " Synchronous alignement", Rapport de Recherche de l'IDIAP, RR-99-06

[Mar97] Martin A. and others. (1997), "The {DET} curve in assessment of detection task performance ", Eurospeech, vol 4, pp. 1895-1898.

[Ros92] Rosenberg A-E. *et al.* (1992), "On the use of cohort normalized score for speaker verification", in Intl. Conf. On Spoken Language Processing, pp. 599-602.

AMELIORATION DU RECUIT SIMULE : APPLICATION AU PROBLEME D'ASSIGNATION D'INDICE

M. BOUZID, B. BOUDRAA, M. BOUDRAA & B. GUERIN
Institut d'électronique, USTHB, BP 32 EL-ALIA, Bab-Ezzouar, ALGER, ALGERIE.
E-mail : mbouzid@yahoo.com

Abstract. *In this paper, we present a variant of the simulated Annealing algorithm (SA) that we have earlier developped. Indeed, a drawback of the classical algorithm is the probabilistic lost of the absolute minimum during the progress of the algorithm. The new variant consists of memorizing the absolute minimum value of the energy encontred during the whole progress of the algorithm. One pass is suffisant and best results are obtained using this new version. A comparative evaluation between the standard-SA algorithm and its improved version is presented.*

Mots clés : Recuit Simulé, Codage Canal-Source, Assignation d'indice.

I. INTRODUCTION

Dans le cas des systèmes à *QV* conçus initialement pour des transmissions à travers un canal non bruité (idéal), le codage canal-source sans redondance par la technique d'assignation d'indice (*AI*) peut fournir une robustesse contre les erreurs de canal, lorsque ces systèmes sont appelés à opérer dans un environnement perturbé [1,2,3,4,5,6]. Ainsi, la distorsion des paramètres peut être considérablement réduite par une assignation judicieuse de mots-codes binaires aux indices des vecteurs-codes de ces systèmes. Un algorithme d'optimisation itératif basé sur le principe du recuit simulé (*RS*) a été développé pour mettre en œuvre ce codage de canal-source par *AI*. Cet algorithme a pour objectif de trouver des codes canal-source (vecteurs d'*AI*) globalement optimaux, destinés à la protection implicite des indices des dictionnaires d'un codeur *CELP* (Code Exited Linear Prediction) fonctionnant 4.8 KBPS [4,7]. Nous avons montré dans [4,5,6] que ce type de codage de canal par *AI*, non redondant, fournit une bonne protection aux indices des dictionnaires du codeur *CELP* contre des erreurs aléatoires de transmission. Ainsi, une amélioration significative des performances objectives du codeur-*CELP* en présence de bruit de canal a été apportée par cette technique, sans avoir recours à des techniques de codage de canal redondant.

Bien que l'algorithme du *RS* soit connu pour son efficacité à résoudre des problèmes d'optimisation combinatoire, il peut cependant présenter l'inconvénient de ne pas garantir une solution globalement optimale (optimum absolu). Ceci est dû à une de ses propriétés qui peut faire perdre à l'algorithme le minimum global durant son déroulement global. Pour remédier à cet inconvenient, nous avons introduit une variante qui consiste à retenir l'optimum absolu parmi les minimums engendrés durant tout le déroulement de l'algorithme.

II. CODAGE CANAL-SOURCE PAR *AI*

Pour un système à *QV* conçu à base d'un dictionnaire de taille $L=2^n$ vecteurs-codes, le vecteur d'*AI* (ou configuration) affecte à chaque vecteur-code d'indice $i \in \{0,1\}^n$ un mot-code binaire unique, appartenant au même ensemble. $\{0,1\}^n$ est l'ensemble des L indices binaires associés aux vecteurs-codes y_i (i = 0,..., L-1) [1,2,4,5] et n représente le nombre de bits par indice binaire. On symbolise ce vecteur d'*AI* par $\boldsymbol{b}$ ($\boldsymbol{b} \equiv (b(0), b(1),..., b(L-1))$, où $b(i)$ est le mot-code attribué à l'indice binaire i, c-à-d., au vecteur-code y_i. Le problème de la recherche du meilleur réarrangement des vecteurs d'un dictionnaire donné, entraîne la recherche parmi toutes les assignations d'indices possibles, de celle qui assure la meilleure performance possible; c-à-d celle qui minimise la distorsion moyenne totale du système de transmission [2,4,5]. Selon certaines suppositions, le problème peut être réduit à la simple minimisation de la distorsion due au bruit de canal ε_C (critère d'erreur). Dans cette étude on considère le cas d'un seul bit d'erreur dominant par mot-code bruité. Ainsi, notre critère d'erreur approprié à l'optimisation des mots-codes est donné d'une manière simplifiée par [4,5,6] :

$$\varepsilon_C = \sum_{i=0}^{L-1} p(i) \sum_{m=1}^{n} d(y_i, y_{b^{-1}(f(b(i),m))}) \qquad (1)$$

où $p(i)$ dénote la probabilité d'occurrence d'un vecteur-code y_i; $d(x,y)$ représente la distorsion ou de distance entre x et y; $f(i, m)$ est l'indice du vecteur-code obtenu par inversion (bruitage) du $m^{\text{ème}}$ bit de l'indice i.

III. *AI* PAR LA METHODE DU *RS*

La recherche du vecteur d'*AI*, $b(.)$, qui minimise le critère d'erreur (1) est un problème d'optimisation combinatoire. Il est souvent impraticable d'évaluer les performances pour toutes les configurations possibles. Aussi, une méthode sous-optimale est utilisée pour résoudre ce problème. Dans ce type d'optimisation, le *RS* s'avère un outil puissant. Cet algorithme à caractère itératif est souvent utilisé pour trouver des solutions à une variété de problèmes d'optimisation combinatoire [8,9]. La méthode du *RS* est fondée sur une analogie entre la recherche d'un état d'énergie minimale (état stable), pour un système physique, et la minimisation d'un critère d'erreur, pour un problème d'optimisation combinatoire [8,9]. Chaque vecteur d'*AI*, $b(.)$ simule un état d'un système physique avec un critère d'erreur ε_C, qui représente l'énergie du système.

IV. AMELIORATION DE L'ALGORITHME DU *RS*

Le *RS* présente l'avantage de ne pas rester bloqué dans des minimums locaux. Ceci grâce à sa propriété qui permet l'acceptation d'une configuration qui fait accroître l'énergie. Cependant, cette propriété peut présenter l'inconvénient de faire perdre à l'algorithme le minimum global et l'acceptation de configurations moins bonnes. Ainsi, plusieurs déroulements de l'algorithme sont alors nécessaires, avec les mêmes valeurs des paramètres de contrôle, afin d'avoir la certitude que la solution trouvée est optimale ou proche de l'optimale. Ceci, va induire une large consommation du temps de traitement; surtout pour les dictionnaires dynamiques. Pour remédier à ce problème, nous avons introduit une modification dans l'algorithme du *RS*. Elle consiste à comparer, à chaque itération, la valeur de l'énergie actuelle à la valeur minimale déjà rencontrée et retenir, finalement, la configuration optimale qui a engendré le minimum absolu. L'algorithme du *RS* utilisant cette approche est nommé algorithme du recuit simulé amélioré (*RSA*). On représente en italique les modifications apportées.

Algorithme du RSA :

Etape 1: *Initialisation :*

- définir un paramètre température T et l'initialiser à une valeur haute : Ti.
- choisir, selon une méthode appropriée, l'état initial b.
- Initialiser : Emin $=\Delta\varepsilon_0$ ($\Delta\varepsilon_0 = \Delta\varepsilon$ de la 1er iteration) *et* $b_{opt} = b_0$ (b_0 *de la 1er iteration)*

Etape 2: *procédure de Metropolis* :

- choisir aléatoirement un état b' : perturbation de l'état b et calculer la variation d'énergie: $\Delta\varepsilon = \varepsilon_C(b') - \varepsilon_C(b)$.
- ***Règle d'acceptation de Metropolis:***

a) si $\Delta\varepsilon < 0$, remplacer b par b', aller à l'étape 3.

b) si $\Delta\varepsilon \geq 0$, remplacer b par b' avec la probabilité : $\exp(-\Delta\varepsilon / T)$, aller à l'étape 3.

- ***Règle de sauvegarde de l'etat d'energie minimal :***

Si l'etat b est accepté et $\Delta\varepsilon <$ Emin :
Emin $=\Delta\varepsilon$ et $b_{opt} = b$

Etape 3: *Test d'équilibre thermodynamique local*

- si dans l'étape 2 le nombre de baisse d'énergie dépasse un nombre prescrit ou si trop de perturbations qui n'engendrent aucune baisse d'énergie, se produisent aller à l'étape 4.
- Sinon, retour à l'étape 2.

Etape 4: *Programme du recuit* :

- diminuer lentement T : $T_k = \alpha\ T_{k-1}$, $0 < \alpha < 1$

Etape 5: *Test de fin* :

- si $T < Tc$ (Tc : température de congélation prescrite) ou s'il apparaît qu'un état stable est obtenu, alors fin du programme avec b_{opt} qui décrira l'état stable du système.
- Sinon, retour à l'étape 2.

A l'étape initiale, on doit choisir une température suffisamment haute pour éviter que le processus du *RS* ne soit piégé, dès le départ, dans un minimum local. Pour l'*AI* initiale, un algorithme basé sur la méthode de l'amélioration itérative a été mis au point pour la recherche de la configuration initiale [4]. Dans l'étape 2, les perturbations b du vecteur d'*AI* sont effectuées en inter-changeant aléatoirement les assignations de deux vecteurs-codes seulement. Le changement dans la configuration est accepté ou rejeté selon la règle d'acceptation de Metroplis. Dans l'étape 2a), $\Delta\varepsilon < 0$ correspond à une diminution de l'énergie. D'où, l'état d'essai est accepté comme prochain état du système avec une probabilité égale à un. D'autre part, $\Delta\varepsilon \geq 0$ correspond à une augmentation de l'énergie. L'état d'essai peut tout de même être accepté avec une probabilité $P(\Delta\varepsilon) = \exp(-\Delta\varepsilon/T)$, de sorte que l'énergie décroît au fur et à mesure que la température diminue. Ce processus permet à l'algorithme de s'échapper des minimums locaux. Pour un palier de température, on effectue une série de perturbations. Lorsque le système atteint son équilibre thermodynamique local, on abaisse légèrement la température et on effectue une nouvelle série de perturbations à température fixe (palier suivant). Il est important de mentionner que les paramètres de contrôle de l'algorithme du *RS* (Ti, α, Tc et autres [4]) sont obtenus expérimentalement, de sorte que les résultats soient satisfaisants et que le temps de calcul reste acceptable.

V. EVALUATION COMPARATIVE DES DES ALGORITHMES : *RS STANDARD* ET *RS AMELIORE*

Des codes de canal dépendant de la source (vecteurs *AI*), obtenus par application de l'algorithme du *RSA*, sont utilisés pour protéger la transmission des indices des dictionnaires d'un codeur *CELP* de 4.8 KBPS. Ce codeur opère sous des conditions spécifiées d'erreurs aléatoires de transmission. L'implantation de ce codeur est effectuée dans les conditions suivantes: L'excitation est modélisée par une combinaison linéaire de deux vecteurs-codes (modélisation d'ordre 2). Ces deux vecteurs sont sélectionnés selon une procédure d'analyse par synthèse. Ceci est effectué en minimisant un critère d'erreur perceptuel [7,4]. Le premier vecteur, d'indice de retard $j(1)$, est extrait d'un dictionnaire prédictif de 128 vecteurs et le second (d'indice statistique $j(2)$) d'un dictionnaire d'excitation statistique de même taille. Les deux vecteurs sont pondérés respectivement par les gains $g_{j(1)}$ et $g_{j(2)}$. Les paramètres transmis au canal, par le codeur *CELP*, sont les coefficients de prédiction LAR_i (rapports d'aires), les indices $j(1)$ et $j(2)$ ainsi que les gains correspondants $g_{j(1)}$ et $g_{j(2)}$. Le codage des coefficients LAR_i et des gains de pondération est réalisé par une méthode de quantification scalaire. Les indices prédictif et statistique $j(1)$ et $j(2)$ sont codés directement en binaire par une *QV* [4]. Le tableau 1 montre l'allocation des bits des paramètres à transmettre par fenêtre d'analyse de 30 ms.

Tableau 1 : Allocation des bits

Paramètres	bits alloués	débit (bits/s)
10 coefficients LAR_i	6,6,5,5,5,5,4,4,4,4	1600
$4 \times j(1) + 4 \times g_{j(1)}$	$4 \times 7 + 4 \times 5$	1600
$4 \times j(2) + 4 \times g_{j(2)}$	$4 \times 7 + 4 \times 5$	1600
Total	144	4800

Les résultats objectifs obtenus sous différentes conditions de transmissions bruitées, ont été jugées satisfaisants. Ceci a confirmé l'efficacité de la technique de codage canal-source adoptée dans la protection implicite des paramètres à transmettre [4,5,6]. Ces résultats ont été obtenus par application de l'algorithme du *RSA*. Ce dernier diffère du *RS standard*, par le fait qu'il fournisse une solution proche de l'optimale au problème d'*AI*, avec les mêmes valeurs des paramètres de contrôle que le *RS*.

Nous donnons à la figure 1 des résultats comparatifs obtenus dans le cas de la transmission des indices statistiques à travers un canal bruité. Le signal utilisé pour la comparaison des performances objectives est d'environ 16 secondes. Il est constitué de 5 séquences de parole, prononcées par 2 locuteurs masculins. Les séquences de parole sont extraites d'un large corpus de phrases arabes phonétiquement équilibrées [10]. Sous des conditions idéales de transmission, la performance objective du codeur *CELP* en terme de *RSBsegmental* moyen est de **10.22 dB**. Les sont presentés à la figure 1. L'analyse des résultats obtenus, pour des taux d'erreurs de transmission variant entre 10^{-1} et 10^{-2}, montre clairement la différence de performance des deux codages de canal dans le cas des indices statistiques. Le codage de canal obtenu par l'algorithme du *RSA* assure ainsi une meilleure performance objective par rapport à celle obtenue pour le codage de canal par du *RS standard*. Ceci pour toute la plage de variation du taux d'erreur de transmission (0.01- 0.1). Notons aussi que pour des taux d'erreurs de transmission inférieurs à 0.02, la différence des performances obtenues par l'application des deux codages de canal est faible, voire même négligeable (cf. figure 1). Cependant, pour des taux d'erreurs relativement élevés la différence de performance est apparente.

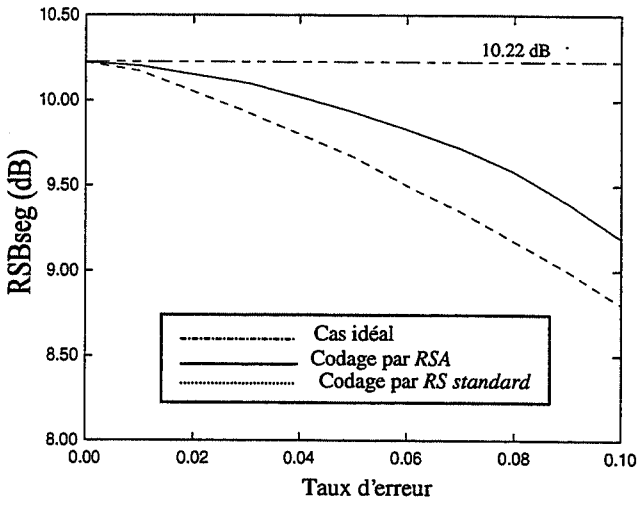


Figure 1. Comparaison des performances entre les deux versions du *RS*

CONCLUSION

Nous pouvons conclure que pour des transmissions à faible taux d'erreur, le codage de canal obtenu par *RS standard* peut assurer les mêmes performances que celles obtenues par le codage de canal issu d'une exécution du *RSA*. Par contre, pour des taux d'erreurs élevés, l'emploi d'un codage de canal obtenu par l'algorithme du *RSA* peut apporter une amélioration significative des performances du codeur *CELP* dans un milieu bruité. Ainsi, cette nouvelle approche a contribué à l'amélioration des performances du codage canal-source-*AI* des indices du dictionnaire-*CELP*.

BIBLIOGRAPHIE

[1] L.C. POTTER & D.M . CHIANG, " Minimax Nonredundant Channel Coding ", IEEE Trans on communications, Vol. 43, N°.2/3/4, pp. 804-811, Feb/ Mar /April 1995.

[2] K. ZEGER & A. GERSHO, " Pseudo Gray Coding", IEEE Trans on Communications, Vol.38, pp.2147-2158, Dec 1990.

[3] D. M. CHIANG & L .C. POTTER, " Graph Covering Index Assignment In Vector Quantisation For Noisy Channel", Appeared in Electronics Letters, Vol. 31, N°.18, Aug ust 1995.

[4] M. BOUZID, " Codage de Canal à Source Dépendante pour des Transmissions par Canaux Bruités. Application au codeur *CELP*", thèse de MAGISTER, Institut d'Electronique, USTHB, Avril 1998.

[5] B. BOUDRAA, M. BOUZID, M. BOUDRAA & B. GUERIN, "Codage de canal à source dépendante : application au codeur *CELP*", XXII^{èmes} JEP, pp. 315-318, Martigny- Suisse,15-19 Juin 1998.

[6] M. BOUZID, B. BOUDRAA & M. BOUDRAA " Assignation d'indice par Recuit Simulé. Application à un Codeur *CELP* de 4.8 KBPS ", JTEA'98, en Additif, Nabeul, Hammamet - Tunisie, 6-7 Novembre 1998.

[7] N.MOREAU, " Codage prédictif du signal de parole à débit réduit: une présentation unifiée", Annales des telecom, Vol. 46, N° 3-4 , 1991.

[8] S. KIRKPATRICK, C. D. GELATT & M. P. VECCHI, "Optimisation by Simulated Annealing", Science, Vol. 220, N°. 4598, pp.671-680, 13 May 1983.

[9] P. SIARRY, " La méthode du recuit simulé : théorie et applications", APII. Vol.29, N°.4 -5, pages 535-561, 1995.

[10] M. BOUDRAA, B. BOUDRAA & B. GUERIN, " Mise en place de phrases arabes phonétiquement équilibrées", XIX^{éme} JEP, Bruxelles, Mai 1992.

Détection de la modulation d'amplitude liée au voisement : comparaison entre expérimentation et modélisation

Angélique GROSGEORGES[1], Frédéric BERTHOMMIER[1]
Frédéric APOUX[2], Christian LORENZI[2]

[1]Institut de la Communication Parlée/INPG
46, Av. Félix Viallet
38031 Grenoble CEDEX
{bertho, ggeorges}@icp.inpg.fr

[2]LPE, UMR CNRS 8581
71, Avenue E. Vaillant
92774 Boulogne-Billancourt
{apoux, lorenzi} @psycho.univ-paris5.fr

ABSTRACT

This study investigates the modelling of perception of the voicing cue and harmonicity detection for speech. The decision model is based on a voicing index extracted after subband autocorrelation and the choice of a threshold. Using the same set of stimuli composed of 16 VCV logatomes as in Lorenzi et al.'s [Lor99] consonant identification task, we carry out a detailed comparison between model's predictions and behavioral responses for the transmission of the voicing cue. The stimuli were spectrally degraded and their envelope were modified in order to vary the modulation depth in four subbands. We reduce the task to a binary decision about the voicing quality of the consonant. After tuning the model's threshold, we show that it is able to account for the 12 normal-hearing listeners' detection performance.

1. INTRODUCTION

De nombreuses études ont mis en évidence le rôle fondamental joué par l'enveloppe temporelle du signal dans le processus de perception et de reconnaissance de la parole. En modifiant les signaux de façon adéquate, ces études montrent que l'information temporelle véhiculée par la modulation d'amplitude (MA) de fréquence inférieure à 500Hz contribue à l'identification de la parole ([Dru94],[Sha94]). L'expérience réalisée par Lorenzi et coll. ([Apo98],[Lor99]) apporte des arguments supplémentaires dans ce sens. Celle-ci est fondée sur un paradigme d'identification de consonnes faisant appel a un signal appauvri spectralement, mais dont l'information temporelle fine est en partie préservée. Lorsque les consonnes sont divisées en deux sous groupes, voisée/non voisée, cette étude montre que ce trait est transmis, c'est à dire qu'il influence les sujets dans une tâche d'identification. Les résultats montrent également que cette transmission dépend de deux facteurs qui modifient la profondeur de MA (niveau de bruit et application d'un exposant sur le signal d'enveloppe temporelle). Par contre, le mode et le lieu d'articulation semblent mal transmis dans les mêmes conditions.

Parallèlement, et en vue d'améliorer les systèmes automatiques de reconnaissance de la parole, une étude a été menée sur l'usage du trait de voisement afin de détecter, marquer et renforcer les régions du plan temps-fréquence où le signal de parole domine un bruit interférant ([Ber98],[Ber00]). Nous explorons le lien existant entre ces deux approches, en montrant que le mécanisme d'extraction auquel nous faisons appel pour la reconnaissance automatique peut sous-tendre les décisions humaines sur le voisement des consonnes. En particulier, ces deux études ont pour hypothèse commune que l'information liée au voisement et à l'harmonicité des signaux de parole est codée temporellement dans le système auditif. Nous faisons donc appel à un indice temporel en sous-bandes similaire à celui que nous utilisons pour le marquage temps-fréquence, afin de construire un modèle décisionnel capable de rendre compte globalement des réponses psychoacoustiques de 12 sujets normo-entendants (l'analyse des données que nous présentons, concernant les facteurs de MA, est détaillée dans [Apo00]).

2. STIMULI

Les stimuli analysés par le modèle sont identiques à ceux utilisés expérimentalement. Un ensemble de 48 signaux dits "clairs " a été construit à partir d'une série de 3 enregistrements de chacun des 16 logatomes de type /aCa/, comprenant 10 C voisées (C=/b,d,g,v,ʒ,z,m,n,r,l/) et 6 C non voisées (C=/p,t,k,f,s,ʃ/), lus par une locutrice française dans le silence. A chacun de ces logatomes un bruit blanc stationnaire gaussien est additionné avec un rapport signal sur bruit (RSB) à 3 niveaux possibles (-6, 0, +6 dB). Ensuite ces signaux (signal clair+bruit) sont traités en sous-bandes de façon à éliminer l'essentiel des caractéristiques spectrales (le peigne harmonique, ainsi que les pics formantiques). Une décomposition fréquentielle en 4 sous-bandes est appliquée, puis l'enveloppe temporelle des signaux est extraite par démodulation, avec un filtre passe-bas de fréquence de coupure égale à 500Hz. La MA générée par le battement des composantes harmoniques dans des sous-bandes larges persiste après ce traitement. Le signal d'enveloppe temporelle obtenu reste à la puissance k=1 dans la condition " non renforcé " ou bien est élevé à la puissance k=$\sqrt{2}$,2 dans les conditions " renforcé ". Cette enveloppe est ensuite utilisée pour moduler un bruit blanc, puis les signaux des 4 sous-bandes sont re-filtrés et additionnés. Finalement, la profondeur de MA des stimuli se trouve contrôlée selon deux facteurs à 3 niveaux: (1) le RSB (-6, 0, +6 dB) (2) l'exposant k du renforcement des enveloppes temporelles (k = 1, $\sqrt{2}$, 2). Ainsi, notre de base données est composée de 9 blocs de 48 stimuli.

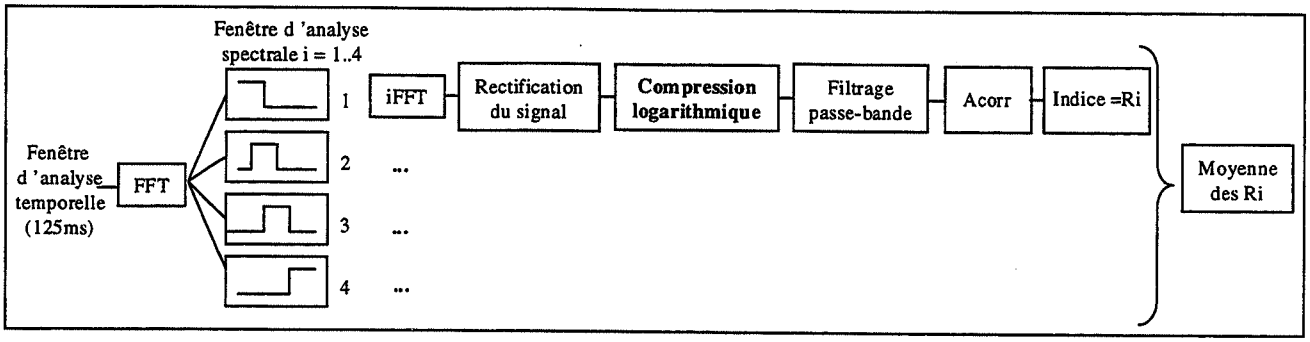


Figure 1: Bloc-diagramme du calcul de l'indice de voisement. L'extraction de l'indice Ri est effectuée pour chaque trame temporelle et dans chacune des sous-bandes i après démodulation. Le signal temporel est rectifié, comprimé logarithmiquement, puis filtré passe-bande dans le domaine de la fréquence fondamentale (f0). Ri est l'amplitude normalisée du pic de l'auto-corrélogramme observé dans le domaine de f0. Pour chaque stimuli, la moyenne spectrale et temporelle de Ri est prise en compte.

3. L'INDICE R

Les signaux sont analysés afin d'associer à chacun des stimuli une mesure qui servira de support à un processus décisionnel de voisement. Le bloc diagramme du processus d'extraction est illustré figure 1 (voir une description plus complète dans [Ber00]). Le modèle utilisé est fondé sur l'analyse temps-fréquence (TF) par banc de filtres auditifs proposée par [Tes99]. Le plan TF est divisé en régions rectangulaires dans lesquelles une mesure est effectuée localement. Temporellement, chaque trame rectangulaire dure 125ms avec un recouvrement de moitié. Spectralement, le signal est décomposé en 4 sous-bandes définies à partir de fenêtres de Hanning disposées selon une échelle Bark.

Localement, pour chacune des régions TF, l'extraction de l'indice nommé Ri est réalisée en effectuant l'autocorrélation du signal. Celui-ci est préalablement compressé logarithmiquement et démodulé. La phase de compression a été rajoutée au processus d'extraction que nous utilisons habituellement pour le calcul de Ri. La démodulation proprement dite s'effectue en rectifiant le signal puis en appliquant un filtrage passe-bande trapézoïdal [0, 90, 350,1000]Hz privilégiant le domaine de la fréquence fondamentale (f0) des signaux de parole. Enfin, la valeur de Ri=R1/R0 est calculée en effectuant le rapport entre R1, l'amplitude premier pic de l'autocorrélogramme trouvé dans l'intervalle de période fondamentale 1/[350,90]s, et R0, l'amplitude à t=0. La moyenne R des indices Ri locaux est calculée pour chaque stimuli.

Les moyennes de R sont établies pour chacune des 9 conditions afin de représenter l'effet des deux facteurs (figure 2). Nous vérifions ainsi que les moyennes de R varient en fonction de la profondeur de MA, que les facteurs RSB et exposant augmentent tous les deux. Un graphique comparable est établi pour les sujets (figure 3), à partir de la quantification proposée par [Apo00] (en % de transmission de l'information de voisement). Nous voyons qu'il est possible de rendre compte de la décision des sujets à partir de l'indice R, à condition de disposer d'un modèle de décision.

Pour construire un modèle de décision ayant une performance significative, une condition à remplir est que les moyennes des deux classes à discriminer soient suffisamment différentes, et que les distributions ne se recouvrent pas. La compression logarithmique s'avère nécessaire pour obtenir cette différenciation dans plusieurs blocs. Pour l'un de ces blocs, la figure 4 montre l'effet indispensable de la compression sur les répartitions, ainsi que sur les moyennes des classes "voisée/non voisée".

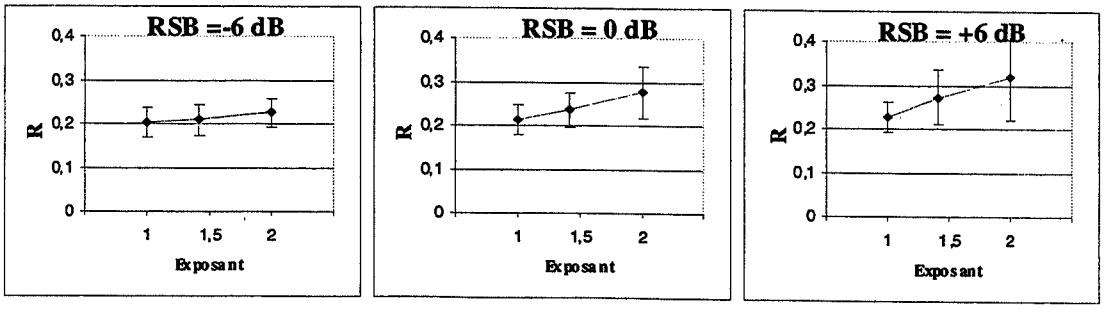


Figure 2: Variation de la moyenne de R par bloc (sur 48 stimuli), en fonction des deux facteurs de contrôle de la MA: RSB et exposant. Les barres d'erreur représentent ± 1 écart-type.

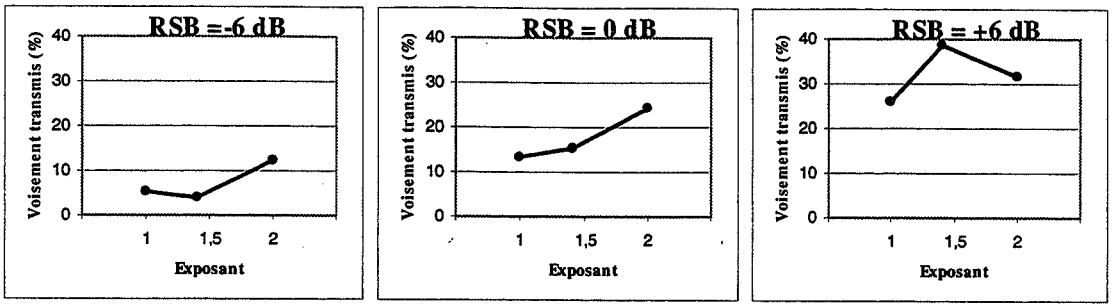


Figure 3: Résultats expérimentaux, exprimés en % d'information transmise sur le voisement dans une tâche d'identification des consonnes (d'après [Apo00]).

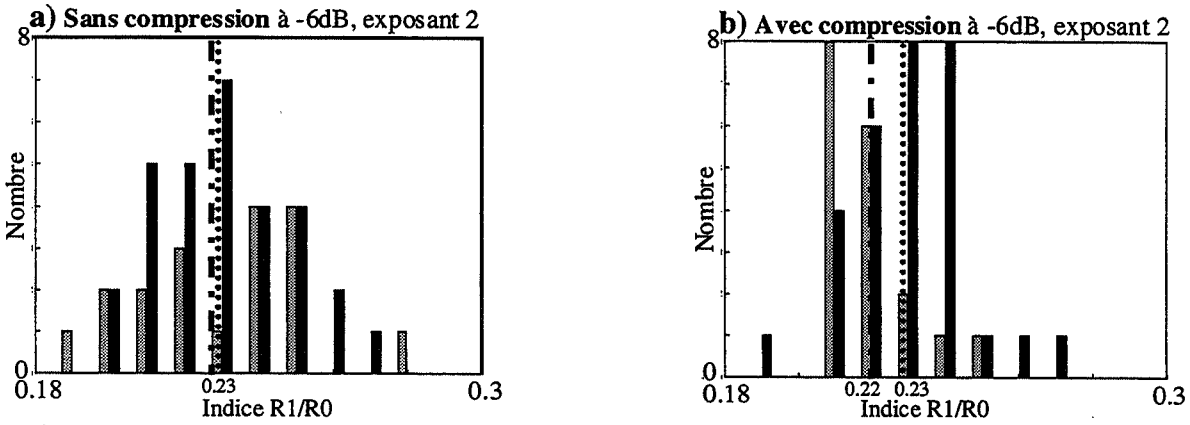


Figure 4: Effet de la compression logarithmique sur les répartitions (sans compression fig.4a; avec compression fig.4b) des indices des 30 stimuli à C voisée (foncé) et des 18 stimuli à C non voisée (clair), dans une condition (-6 dB RSB, exposant 2). Avec la compression, nous observons une différenciation des moyennes (voisée en pointillé, non voisée en tiret-point).

4. MODELE DE DECISION

La construction d'un modèle de décision requiert la fixation d'un seuil de détection de voisement. Nous fixons ce dernier de deux manières différentes en appliquant deux contraintes, l'une intrinsèque et l'autre pour simuler la réponse des sujets à partir du modèle.

Première méthode: On ne tient compte que des statistiques d'estimation de l'indice. Pour chaque bloc, nous fixons le seuil à partir des distributions de l'indice R observées pour les stimuli répartis en deux groupes: C voisée (30) et C non voisée (18). Le critère que nous optimisons est la somme sensibilité (Se) + spécificité (Sp) du détecteur. Un exemple est montré figure 5, où nous traçons la fonction Se=f(1-Sp) (courbe ROC) correspondant au déplacement du seuil de 0 à 1. Nous figurons par une croix le point répondant à la maximisation de la somme (Se+Sp). Celui-ci est lié à une valeur de seuil optimale selon ce critère. Nous obtenons ainsi le "seuil du modèle".

Seconde méthode: Nous fixons le "seuil des sujets" en partant de leur matrice de décision (table 1a), ainsi que des distributions d'estimation de l'indice. Pour chacun des 9 blocs, nous recherchons le seuil des sujets en minimisant la différence entre la matrice de décision du modèle obtenue pour un seuil donné et la matrice de décision des sujets (qui est fixe). Cette recherche est effectuée en déplaçant le seuil comme précédemment, puis en sélectionnant le paramètre de seuil du modèle correspondant au point de la courbe Se=f(1-Sp) le plus proche du point (1-Sp, Se) obtenu à partir de la matrice de décision des sujets (figure 5, triangle). Ainsi, il est possible de fixer un seuil à partir des réponses modélisées bloc par bloc pour simuler globalement les réponses des sujets. Pour un bloc, la matrice de décision du modèle correspondante à ce seuil est établie pour comparaison (table 1b).

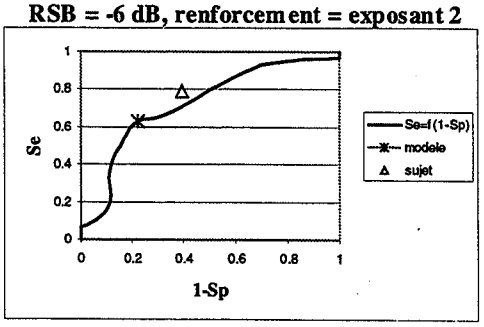


Figure 5: Courbe ROC du modèle pour 1 bloc. Procédure de détermination des seuils à partir des mesures de l'indice Ri sur le signal. 1ère méthode: Se et Sp sont calculées pour chaque valeur de seuil. Le seuil établi est associé au point max(Se+Sp), figuré par une croix sur ce graphique. 2ème méthode: le triangle représente le point extrait de la matrice de décision des sujets, qui est ensuite associé au seuil de décision des sujets (voir texte).

Table 1: Matrice de décision obtenue avec la seconde méthode pour la condition : exposant k=2, RSB = –6 dB.

a) Matrice expérimentale de décision des sujets

Stimuli C	Voisé	Non voisé
Voisé	Se =79 %	21 %
Non voisé	1-Sp=39 %	61 %

b) Matrice de décision du modèle au seuil des sujets

Stimuli C	Voisé	Non voisé
Voisé	Se=76 %	24 %
Non voisé	1-Sp=38 %	62 %

Comparaison: Les couples de seuils obtenus avec les deux méthodes, pour chacun des 9 blocs, sont compilés dans la table 2, où nous voyons qu'ils sont quasiment identiques. Enfin, nous comparons les performances (Se+Sp) des sujets et celles du modèle ajustées sous la contrainte de maximisation de (Se+Sp) (table 3).

Table 2: Seuils de décision pour les 9 blocs.
Seuil de décision des sujets/ *seuil de décision du modèle*

	RSB		
Exposant	**-6**	**0**	**6**
1	0.20 / *0.19*	0.21 / *0.21*	0.22 / *0.22*
1.41	0.21 / *0.20*	0.23 / *0.23*	0.26 / *0.25*
2	0.21 / *0.22*	0.26 / *0.27*	0.30 / *0.31*

Table 3: Performances (Se+Sp) pour les 9 blocs.
(Se+Sp) des sujets/ *(Se+Sp) du modèle*

	RSB		
Exposant	**-6**	**0**	**6**
1	1,21 / *1,32*	1,44 / *1,48*	1,45 / *1,52*
1.41	1,34 / *1,29*	1,50 / *1,61*	1,69 / *1,71*
2	1,38 / *1,41*	1,63 / *1,64*	1,73 / *1,80*

5. CONCLUSION

Les différences de seuil de décision sont très faibles, et les différences de performance sont suffisamment réduites pour conclure **(1)** que le modèle peut rendre compte des réponses psychophysiques bloc par bloc, et **(2)** que la contrainte de maximisation de (Se+Sp), choisie pour fixer le seuil du modèle est proche du critère qui sous tend la réponse des sujets.

Cette conclusion accorde un degré de plausibilité au processus d'analyse que nous proposons. Celui-ci possède des points communs avec des modèles classiques d'analyse de la MA, tel que le détecteur linéaire d'enveloppe de Viemeister [Vie79], en particulier le mode d'analyse temporel. Une méthode de détection de voisement fondée sur une analyse spectrale est inapplicable sur nos stimuli. Par contre, une différence importante avec d'autres algorithmes d'analyse temporelle est liée à l'usage de l'indice R1/R0, et une étude comparative est nécessaire avant de conclure à une supériorité dans notre domaine d'application. Enfin, deux points restent mal expliqués dans cette étude: **(1)** la variabilité du seuil en fonction des conditions, qui interdit à priori la fixation d'un seuil absolu **(2)** l'utilité de la compression logarithmique. Nous pensons que ces deux problèmes seront résolus en améliorant l'étape d'intégration des évaluations locales de l'indice, qui est ici une simple moyenne globale par stimuli. Lorsque les sujets entendent les logatomes /aCa/, la consonne est encadrée par deux segments voisés qui offrent une référence stable pour la décision de voisement.

Remerciements: Ce travail est soutenu par le contrat EEC LTR RESPITE.

BIBLIOGRAPHIE

[Apo98] Apoux, F., Berthommier, F., Bacri, N., Lorenzi, C. (1998) Effet du renforcement des modulations temporelles en sous-bandes sur la reconnaissance de la parole: Résultats préliminaires, JEP, Martigny, pp. 259-262.

[Apo00] Apoux, F., Bacri, N., Berthommier, F., Lorenzi, C. (2000) Effets du renforcement de l'enveloppe temporelle sur l'identification des consonnes dans le bruit, CFA Lausanne, (soumis).

[Ber98] Berthommier, F., Glotin, H., Tessier, E., Bourlard, H. (1998) Interfacing of CASA and partial recognition based on a multistream technique, ICSLP'98, Sydney, pp.1415-1418.

[Ber00] Berthommier, F., Glotin, H. (2000) Reconnaissance de la parole dans le bruit après renforcement fondé sur l'harmonicité, JEP 2000 (ce Vol.).

[Dru94] Drullman, R., Festen, J. M., Plomb, R. (1994) Effect of temporal envelope smearing on speech reception, J. Acoust. Soc. Am., Vol 95, pp. 1053-1064.

[Lor99] Lorenzi, C., Berthommier, F., Apoux, F., Bacri, N. (1999) Effects of envelope expansion on speech recognition, Hear. Res., Vol 136 (1-2), pp. 131-138.

[Sha95] Shannon, R., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M. (1995) Speech recognition with primarily temporal cues, Science, Vol 270, pp. 303-304.

[Tes99] Tessier, E., Berthommier, F., Glotin, H., Choi, S. (1998) A casa front-end using the localisation cue for segregation and then cocktail-party speech recognition, ICSP'99, Seoul, pp. 97-102.

[Vie79] Viemeister, N.F. (1979) Temporal modulation transfer function based upon modulation thresholds, J. Acoust Soc. Am., Vol 66, pp.1364-1380.

Identification Automatique des Langues : variations sur les multigrammes

Jérôme Farinas, Régine André-Obrecht

IRIT - équipe IHM-PT
118, route de Narbonne – F-31062 Toulouse Cedex 04, France
Tél.: ++33 (0)5 61 55 88 35 - Fax: ++33 (0)5 61 55 62 58
Mél: {Jerome.Farinas, obrecht}@irit.fr - http://www.irit.fr

ABSTRACT

Most systems of Automatic Language Identification give a great importance to the phonotactic level, by using N-gram models and relatively large phone-dictionary sizes. However, it is obvious that introducing other features (acoustic, phonetic, prosodic) will improve performances. Recently, we have proposed an alternative acoustic phonetic model which exploits the vowel / non vowel distinction. Here we complete this preliminary system, by studying the phonotactical level and adapting it to the acoustic outputs (small phone dictionary). We used a n-multigram model based on broad phonetic categories. We present a first study based on hand-label data, showing the influence of the number of phonetic broad categories in an ALI task.

1. INTRODUCTION

Parmi les différentes sources d'information disponibles pour identifier un langage donné, les informations phonotactiques, relatives aux règles qui gouvernent la combinaison des sons dans une langues, contribuent grandement à la décision d'identification [Haz97]. Les systèmes actuels les plus performants en témoignent largement en privilégiant cette source de connaissances et sa modélisation [Mat99].

Les études menées à l'IRIT en Identification Automatique des Langues, ont pour but d'exploiter le maximum de sources. C'est pourquoi nous portons nos efforts sur la modélisation acoustico-phonétique, la modélisation phonotactique, la modélisation prosodique et la fusion de ces informations. Une première étude nous a conduit à proposer une approche différenciée au niveau acoustico-phonétique. Pour prendre en compte les paramètres structuraux des systèmes phonologiques, deux espaces, l'espace vocalique et l'espace consonantique, sont modélisés par deux modèles distincts pour chacune des langues. L'identification est obtenue par fusion adéquate des scores ainsi obtenus [Pel00] . Cette approche a montré une amélioration des résultats comparativement à une modélisation acoustique globale. Cette approche remet en cause la modélisation phonotactique. L'influence d'une telle séparation en classes phonétiques sur une modélisation phonotactique est étudiée. Ce premier travail se place dans un cadre idéal dans la mesure où nous utilisons en entrée de la modélisation un étiquetage manuel réalisé par des experts phonéticiens. Nous étudions les performances en reconnaissance de la langue d'un modèle phonotactique à base de multigrammes sur des classes de phonèmes. Nous présentons le modèle multigramme utilisé en section 2, le protocole expérimental en section 3 et nous discutons les résultats en section 4.

2. MODÈLES MULTIGRAMMES

Pour rendre compte des différentes règles qui gouvernent la combinaison des phonèmes d'une langue. nous utilisons un modèle de langage multigramme [Del96] qui permet de détecter des motifs récurrents dans des suites d'observations. Ces motifs récurrents peuvent avoir une longueur variable.

La modélisation par multigrammes consiste à trouver la segmentation $S=(s_1,\dots, s_{n(S)})$ la plus probable d'une séquence d'observations $O=(o_1,\dots, o_T)$:

$$S^* = \arg\max\ \ell(O,S)$$

avec la vraisemblance :

$$\ell(O,S) = \prod_{i=1}^{n(S)} P(z_i)$$

où $Z_i = (o_{s_i}\dots, o_{s(i+1)-1})$

L'algorithme d'apprentissage est un algorithme itératif de type EM. A chaque itération, sont estimées les probabilités *a priori* d'une séquence d'observations Z_i :

$$P^{(k+1)}(z_i) = \frac{c(z_i / S^{*(k)})}{c(S^{*(k)})}$$

avec

$$S^{*(k)} = \arg\max_{S}\ \ell^{(k)}(O,S)$$

où $\ell^{(k)}(O,S)$ est la vraisemblance de la séquence d'apprentissage O à l'itération k, $c(z_i / S^{*(k)})$ est le nombre d'occurrences de z_i dans la segmentation

optimale $S^{*(k)}$. La segmentation la plus probable $S^{*(k)}$ est estimée en utilisant un algorithme de Viterbi. Au cours de ces itérations, les segmentations du corpus évoluent, faisant émerger les séquences d'observation les plus typiques.

Après apprentissage, un dictionnaire est créé contenant les séquences Z_i les plus probables et leur vraisemblance.

La phase de reconnaissance consiste à calculer la perplexité d'une séquence d'observation O en utilisant la segmentation la plus vraisemblable, suivant la formule :

$$PP_{VI}(O) = 2^{-\frac{1}{T}\log \ell^*(O)}$$

où T est le nombre d'observations de O et où la vraisemblance de cette même suite est :

$$\ell^*(O) = \arg \max_S \ \ell(O,S)$$

3. Expériences

Les expériences sont menées sur six langues du corpus OGI Multi Language Telephone Speech : l'anglais, l'allemand, l'hindous, le japonais, le mandarin et l'espagnol. Les données utilisées correspondent aux transcriptions phonétiques réalisées manuellement par des experts phonéticiens [Lan97]. Ces transcriptions, réalisées au format international Worldbet [Hie93], sont ensuite réduites en grandes classes phonétiques. Les voyelles sont regroupées en 9 classes, correspondant à une discrétisation de l'espace articulatoire, suivant les deux premiers formants F1 et F2. Les consonnes sont rassemblées en grandes classes : occlusives (en différenciant le silence avant explosion et l'explosion-friction), fricatives, nasales, liquides et semi-consonnes. De plus, l'information sur le voisement est conservée pour les occlusives, fricatives et nasales. Il en résulte 9 classes consonantiques. Les étiquettes des diacritiques ne sont pas considérées, seules les pauses sont conservées.

Le corpus est scindé en deux parties, l'une destinée à être utilisée pendant la phase d'apprentissage et l'autre pendant la phase de test. La partie destinée à l'apprentissage est constituée d'environ 70 locuteurs par langue, et celle destinée aux tests, 20 locuteurs. Les deux parties sont indépendantes, on ne retrouve pas de locuteur commun entre les deux sous corpus.

Les expériences consistent à faire varier à la fois la longueur maximum autorisée des séquences du modèle multigramme (de 3 à 5) et la composition des classes phonétiques. La variation des classes phonétiques consiste à réduire le nombre de classes des consonnes en regroupant les classes voisées/non voisées, en ne conservant qu'une classe pour les occlusives, en regroupant les sonantes (tableau 1) et en ne considérant qu'un seul des deux axes formantiques pour les voyelles (tableau 2).

Tableau 1 : Description des différents jeux de réduction du nombre de classes pour les consonnes. La dernière colonne indique le nombre de classes consonantiques obtenues après réduction.

jeu	description	nb
#C1	une seule classe consonne	1
#C2	sans distinction de voisement, regroupement des occlusives, regroupement des sonantes	3
#C3	sans distinction de voisement, regroupement des sonantes	4
#C4	sans distinction de voisement, regroupement des occlusives	5
#C5	sans distinction de voisement	6
#C6	regroupement des occlusives, regroupement des sonantes	6
#C7	regroupement des sonantes	7
#C8	regroupement occlusives	8
#C9	aucune réduction	9

Tableau 2 : Description des différents jeux de réduction du nombre de classes pour les voyelles. La dernière colonne indique le nombre de classes vocaliques obtenues après réduction.

jeu	description	nb
#V1	une seule classe de voyelles	1
#V2	axe formantique F1 conservé	3
#V3	axe formantique F2 conservé	3
#V4	sans réduction	9

Pour chaque langue, un modèle phonotactique multigramme est appris comme indiqué ci-dessus. Le problème d'identification consiste alors à trouver la langue qui maximise la probabilité d'observation de la séquence O :

$$L^* = \arg \max_L \ P(O/L)$$

ce qui revient à déterminer :

$$L^* = \arg \max_L \ PP_{VI}(O/L)$$

4. Résultats

Les résultats d'identification correcte en utilisant un modèle 3-multigramme (tableau 3) varient de 33% à 100% pour une tâche de reconnaissance sur 6 langues.

Tableau 3 : Taux d'identification correcte (%) avec un modèle 3-multigramme pour 6 langues

	#V1	#V2	#V3	#V4
#C1	32,9	56,5	60,1	85,6
#C2	53,1	75,3	81,6	93,2
#C3	69,1	90,9	83,1	97,4
#C4	64,5	86,9	84,1	96,2
#C5	75,5	95,1	88,1	98,7
#C6	61,6	84,8	86,4	97,2
#C7	81,6	96,4	88,8	98,7
#C8	71,0	90,9	89,0	98,9
#C9	84,8	97,2	93,6	100,0

Globalement le fait d'autoriser une longueur maximum de 4 observations pour une séquence au lieu de 3 améliore légèrement les résultats (tableau 4).

Tableau 4 : Taux d'identification correcte (%) avec un modèle 4-multigramme pour 6 langues

	#V1	#V2	#V3	#V4
#C1	34,3	63,7	71,0	86,7
#C2	54,0	81,2	84,4	95,5
#C3	72,1	92,6	87,7	97,9
#C4	71,0	84,2	87,9	96,4
#C5	81,2	96,4	90,0	98,8
#C6	70,2	88,8	90,7	98,5
#C7	85,8	96,4	92,6	99,8
#C8	75,9	92,4	93,6	99,1
#C9	88,8	98,3	94,5	100,0

Par contre, en utilisant des 5-multigrammes, même si l'on conserve des résultats proches des 4-multigrammes, les performances se dégradent (tableau 5). La cause essentielle est certainement liée à la taille insuffisante du corpus d'apprentissage pour apprendre de tels modèles : les dictionnaires pour les 5-multigrammes sont alors en moyenne constitués de 1600 séquences d'observations, au lieu de 1000 pour les 3-multigrammes et 2000 pour les 4-multigrammes.

Tableau 5 : Taux d'identification correcte (%) avec un modèle 5-multigramme pour 6 langues

	#V1	#V2	#V3	#V4
#C1	36,9	68,3	68,3	87,9
#C2	63,7	83,5	86,3	94,5
#C3	76,3	92,0	90,3	98,3
#C4	74,0	89,0	88,1	95,7
#C5	81,6	95,7	90,2	99,3
#C6	72,6	91,3	93,0	99,2
#C7	86,3	97,3	94,9	99,6
#C8	74,8	94,0	94,3	99,5
#C9	89,8	98,3	95,5	100,0

Afin d'interpréter plus justement ces résultats, il convient de préciser la répartition des grandes classes phonétiques parmi les langues (tableau 6) : elle est relativement homogène, les voyelles (avec environ 23000 occurrences sur le corpus d'apprentissage) représentent la plus grande partie des occurrences, deux fois plus que les occlusives et les fricatives (resp. 9200 et 9700). Notons cependant l'absence de liquides pour le japonais ; ce biais disparaît dès que les sonantes sont regroupées.

Si l'on compare les résultats obtenus pour les classes où l'on ne prend pas en compte le voisement (ensembles #C2, #C3, #C4, #C5) à ceux pour lesquels on distingue au sein d'une même classe de sons, les segments voisés et non voisés (resp. ensembles #C6, #C7, #C8, #C9), on constate une dégradation d'environ 10% des taux. En effet, certaines langues (hindi, mandarin et espagnol) privilégient des occlusives non voisées dans les séquences les plus probables, alors que d'autres (anglais, allemand) privilégient les occlusives voisées. A noter que le japonais met en avant des séquences d'occlusives avec un voisement mixte : un silence avant explosion non voisé avec une explosion et un relâchement voisé et vice versa.

Si l'on s'intéresse plus particulièrement à la réduction des consonnes liquides, nasales et semi-consonnes en une seule classe, les consonnes sonantes (ensembles #C2, #C3, #C6, #C7 par rapport à #C4, #C5, #C8, #C9), on ne note pas une dégradation des résultats de manière extrêmement sensible : nous ne perdons pas énormément d'information en regroupant ces trois classes phonétiques.

Si nous examinons les dictionnaires n-multigrammes (n=3,4,5) de chaque langue, le score relativement bas obtenu en utilisant une seule classe pour les consonnes et une seule classe pour les voyelles s'expliquent par le fait que les cohortes les plus fréquentes, à savoir CCC, CVC et VCC sont communes à toutes les langues. La cohorte CCC correspond généralement à l'enchaînement d'une fricative ou d'une sonante (C) et d'une occlusive (caractérisée par CC du fait de la distinction entre le silence avant explosion et l'explosion-friction).

Si nous examinons maintenant les séquences les plus fréquentes dans le cas où nous avons le maximum de classes (9 voyelles et 9 consonnes), les séquences les plus fréquentes sont constituées uniquement d'occlusives (anglais, hindou), ou d'occlusives suivi d'une voyelle (allemand, japonais, espagnol), ou d'occlusives suivi d'une fricative (mandarin). Les occlusives se retrouvent la plupart du temps dans les séquences les plus fréquentes.

Tableau 6 : Occurrences des grandes classes phonétiques par langue dans le corpus d'apprentissage

	anglais	allemand	hindou	japonais	mandarin	espagnol
voyelles	21355	23048	24430	23965	20522	28156
occlusives	7182	7521	9808	10160	11107	9580
fricatives	10990	11413	10422	8448	8303	8819
liquides	3228	2635	4281	9	3118	4278
nasales	8562	8355	7258	7394	8562	8482
semi-cons.	3126	1603	3058	3245	3413	3193

5. CONCLUSION

La modélisation multigramme se montre au travers de cette série d'expériences fort appropriée pour rendre compte des règles phonotactiques élémentaires : un modèle 4-multigramme défini sur un ensemble global de 12 symboles comprenant 3 classes consonnes et 9 classes voyelles se montre très performant, tout en utilisant une discrimination (occlusives, fricatives, sonantes et voyelles) qui pourra être effectuée assez aisément de manière automatique. Etant donné que ces résultats sont obtenus à partir d'une classification grossière, il s'agit maintenant de prolonger le modèle phonétique différencié consonne/voyelle, à ces classes de sons afin d'une part de définir automatiquement les symboles phonétiques en entrée du modèle phonotactique et d'autre part de fusionner les scores issus des deux niveaux. Nous envisagerons ensuite la possibilité de fusionner des scores obtenus en utilisant une modélisation prosodique, pour compléter l'utilisation de l'éventail de sources d'information disponibles pour discriminer les langues.

RÉFÉRENCES

[Haz97] T. J. Hazen, & V. W. Zue, (1997), "Segment-based automatic language identification", Journal of the Acoustical Society of America, Vol. 101, No. 4, pp. 2323-2331.

[Mat99] Matrouf D. et al (1999), "Comparing different model configuration for language identification using a phonotactic approach", Eurospeech'99, Budapest, Hongrie, pp 387-390.

[Pel00] Pellegrino F. et al. (2000), "Identification automatique des langues par une modélisation diférenciée des systèmes vocaliques et consonantiques", Reconnaissances des Formes et Intelligence Artificielle, Paris.

[Del96] Deligne S. (1996), Modèles de séquence de longueur variables : application au traitemant du langage écrit et de la parole, Thèse de 3ème cycle, Ecole Nationale Supérieure des Télécommunications, Paris.

[Lan97] Lander T. (1997), The CSLU Labeling Guide, rapport interne, Center for Spoken Language Understanding, Oregon Graduate Institute.

[Hie93] Hieronymous J. L. (1993), Ascii phonetic symbols for the world's languages: WolrdBet. rapport interne, Bell Labs.

Identification des parlers espagnols et détermination expérimentale des indices acoustiques distinctifs

Brigitte Rose

Laboratoire de phonologie de l'U.L.B (Belgique)
Mail : brigitte.rose@usa.net

ABSTRACT

This paper deals with the identification of Spanish vernaculars based on the analyses of a text reading by different speakers. We chose 10 varieties of Spanish from Spain, Central and South America. Speech data were obtained by recording 10 native Spanish speakers from 8 different countries reading a literary text. 1 minute of speech was then extracted from each reading in order to make up the stimulus material for a perception test. According to the results of the experiment, native Spanish speakers are able to identify the main varieties of Spanish on the basis of segmental and prosodic cues. This experiment constitutes the first stage in the determination of a set of reliable cues for the Automatic Identification of Spanish dialects.

1. INTRODUCTION

Les parlers espagnols présentent un large éventail de variations par rapport à la norme sans pour autant que la compréhension entre leurs locuteurs soit menacée. Le but de cet article est de démontrer la capacité de locuteurs provenant de différentes régions d'Espagne et d'Amérique latine à identifier l'origine géographique d'un locuteur donné. Le but final du projet de recherche est de déterminer une série d'indices acoustiques pertinents pour l'Identification Automatique des parlers espagnols, sur le modèle de ce qui a été réalisé pour les parlers arabes par Barkat [Bar 99].

2. GEOGRAPHIE LINGUISTIQUE DU MONDE HISPANIQUE

Tous les parlers espagnols partagent un nombre de traits linguistiques suffisant pour assurer l'unité de la langue. Ils présentent néanmoins de nombreuses variations (phonético-phonologiques, sémantiques et syntaxiques) les uns par rapport aux autres. A partir de ces traits, les dialectologues ont souvent proposé la catégorisation suivante : Espagnol Péninsulaire / Espagnol d'Amérique [Lip 94]. L'identité de nombreux traits phonétiques d'Andalousie, des Canaries et d'Amérique latine (yéismo, seseo, aspiration de /s/, neutralisation de /r/ et /l/ « implosifs »[1] et finaux) a incité d'autres spécialistes à opposer « l'Espagnol atlantique » [Lap 85]. (Andalousie, Canaries et Amérique latine) à l'Espagnol du reste de la Péninsule.

3. CORPUS LINGUISTIQUE ET METHODOLOGIE

Pour représenter les zones géographiques mentionnées, nous avons choisi les 10 variétés linguistiques suivantes:Argentine, Paraguay, Nord et centre du Chili, Colombie, Venezuela (pour l'Amérique du Sud), Mexique (pour l'Amérique centrale), Castille et Andalousie (pour l'Espagne). Une base de données acoustiques a été constituée à partir de l'enregistrement de 10 locuteurs lisant un texte littéraire.

Les enregistrements ont été digitalisés à 44,1 kHz, 16 bits, stéréo. Une minute de parole de chaque locuteur a été sélectionnée et présentée à 40 sujets (10 andalous, 10 madrilènes, 10 chiliens, 10 mexicains) en tant que stimuli pour un test perceptif. Les sujets ont été priés :
1) d'identifier le stimulus en termes de continent ou de partie de continent(Europe/Amérique centrale/du Sud).
2) d'associer le stimulus avec un pays ou une région (Chili /Colombie /Andalousie etc.).
3) de spécifier les indices segmentaux et supra-segmentaux qui ont permis l'identification du stimulus.
En plus de l'identification des indices acoustiques, nous espérons vérifier les hypothèses suivantes :
1) confrontés à un stimulus donné ,les sujets seront capables d'identifier la zone géographique d'où il provient .
2) à l'intérieur de ces zones, les meilleurs taux d'identification seront obtenus pour les parlers les plus proches de la variété du sujet.
3) les identifications erronées concerneront uniquement des parlers de la même zone [Bar 99].

4. RESULTATS

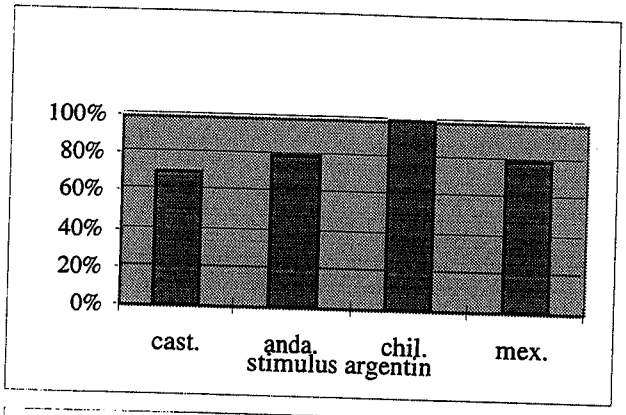
4.1. Identification des parlers espagnols

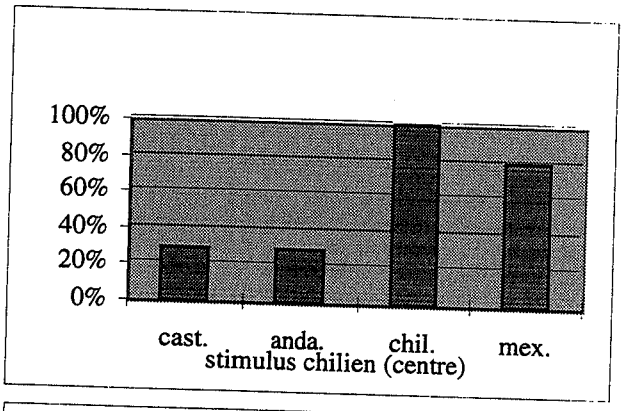
4.1.1. Identification par zones principales : 98,5% des stimuli d'Amérique latine, 95% des stimuli d'Espagne et 100% des stimuli de l'Espagnol atlantique ont été identifiés comme tels. Cela montre l'intérêt de la catégorie « Espagnol atlantique ». 80% des stimuli d'Amérique du Sud et 62,5 % de ceux d'Amérique centrale ont été identifiés correctement.

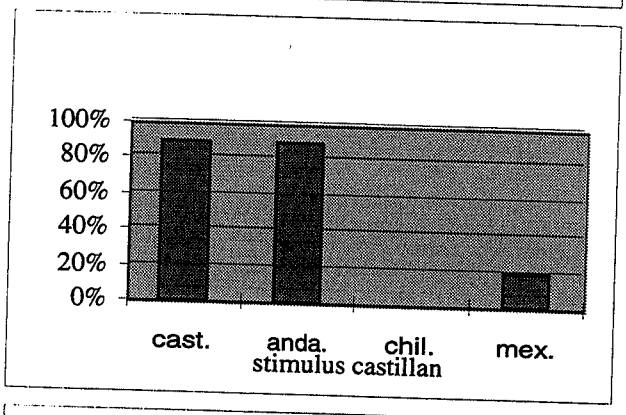
4.1.2. Identification par pays ou par région : Les tableaux suivants reprennent les taux d'identifications correctes par pays (régions pour l'Espagne) selon l'origine

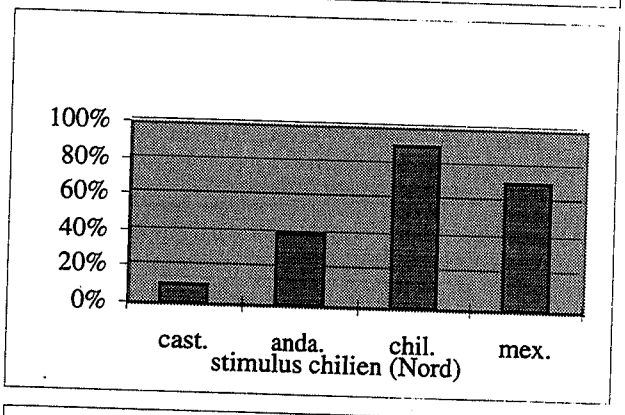
[1] Nous utilisons le terme « implosif » pour désigner les consonnes en position de coda syllabique [Lip 94].

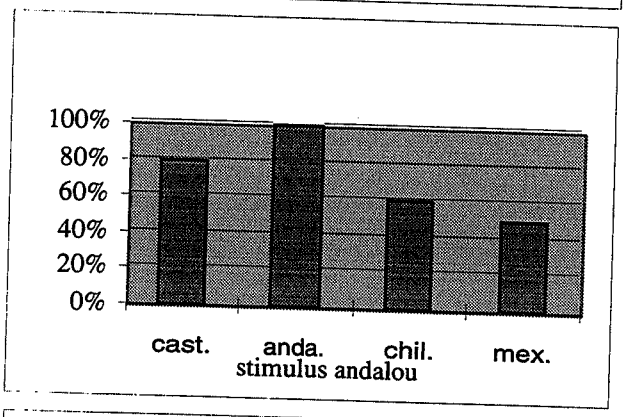
des sujets. Aucun de nos informateurs n'est parvenu à identifier le stimulus paraguayen comme tel.

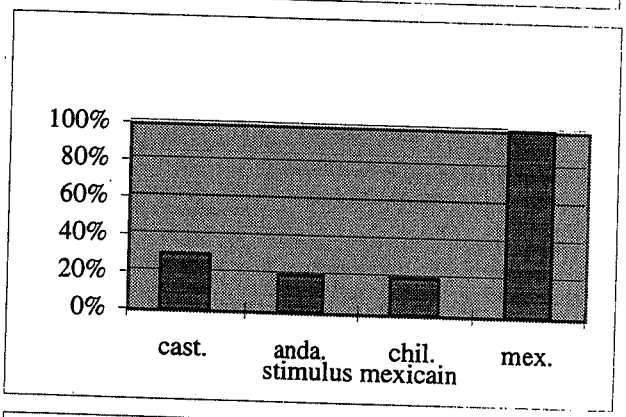


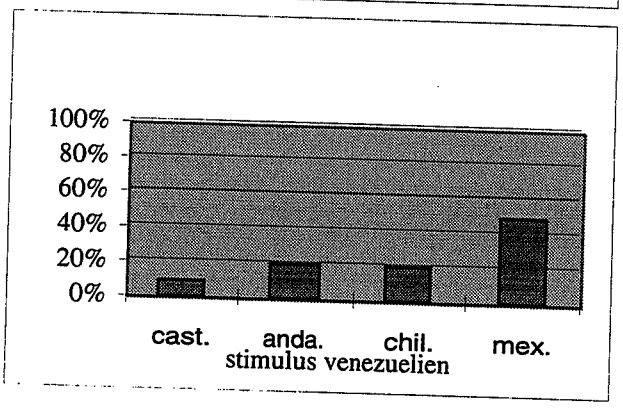


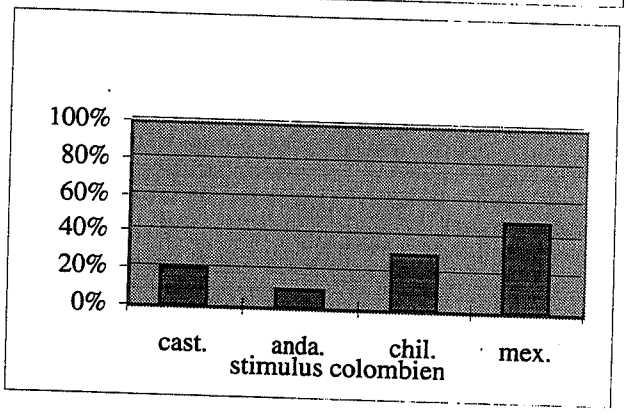


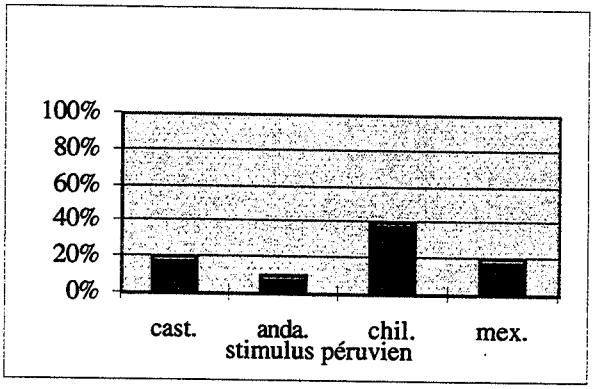












Nos résultats montrent les meilleurs taux d'identification par les sujets :

a) pour leur propre variété d'Espagnol

b) pour les variétés de la même zone que la leur.

4.2. Analyse des erreurs

La plupart des erreurs concernent des variétés de la même zone : confusions entre Chili, Argentine et Uruguay qui ont en effet des traits communs. (/-s/>[h]; /n/ final alvéolaire, /x/ fricative vélaire; /j/>[ʒ]). Néanmoins le taux de confusion entre stimulus d'Amérique centrale et stimulus d'Amérique du Sud s'élève à 20,9%. Cela s'explique par la communauté de traits entre Venezuela, Colombie, Cuba, Saint Domingue et Puerto-Rico (/s/ « implosif » >[h]/ [ʰ]; /x/ >[h]; /n/ final>[ŋ]). Reste à justifier les 18 identifications du stimulus Paraguayen comme Mexicain : malgré certains traits communs (/j/ affriqué ; /n/ alvéolaire) des différences importantes opposent les 2 pays (consonantisme fort du Mexique / faible au Paraguay ; yéismo au Mexique / distinction /j/ et/ʎ/ au Paraguay). Sans doute les racines de l'erreur sont elles à chercher du coté de l'intonation de notre locuteur considérée comme critère premier par les sujets.

4.3. Détermination expérimentale d'indices distinctifs

Les sujets reconnaissent les traits d'un autre parler espagnol à travers le filtre de leur propre variété. Ces traits peuvent être considérés comme a priori pertinents pour l'identification de différents parlers espagnols puisqu'ils sont tenus pour discriminants par les locuteurs hispaniques. Toutefois ces traits doivent être regardés comme représentatifs de la variété régionale de nos informateurs et non de toutes les modalités de leur pays.

4.3.1. Indices phonétiques mentionnés par les sujets

- différentes réalisations de la fricative sourde /s/

indices	pays	exemples
/s/ et/θ/> [s]	Paraguay, Colombie, Chili, Pérou,Mexique	[pjesa] : « pièce» [aβeses] : «parfois»
/s/et/θ/ > [θ]	Argentine, Venezuela, Andalousie	[pjeθa] [aβeθeθ]
/s/ « implosif » >[h]	Venezuela, Chili, Andalousie	[ehpera]: «attente»
/s/final disparait	Para., Argen., Venez., Chili, Anda., Colombie.	[semexante]: "pareils"
/s/ fort	Pérou, Mexique	[semexantes]: "pareils"

Traitement des palatales /j/ et /ʎ/

indice	pays	exemples
/j/≠/ʎ/	Paraguay	[jo]: «moi» [ʎeɣaδo]: «arrivé»
/j/et/ʎ/>/j/	Castille, Venez, Colombie, Anda, Mexique	[jo] [jeɣaδo]
/j/et/ʎ/>[ʒ]	Para, Argen, Chili	[ʒeɣaδo] [ʒo]
/j/et/ʎ/>[dʒ] en position initiale	Argentine, Chili	[dʒo]

- différentes réalisations de la fricative vélaire /x/

indice	pays	exemples
/x/>[x]	Espagne	[axeno]: «étranger»
/x/>[h]	Venezuela, Colombie, Andalousie	[aheno]
/x/+e,i>[ç]	Chili	[açeno]

- différentes réalisations des liquides /r/ et /l /

indice	pays	exemples
/r̄/>[ř]	Paraguay, Chili	[deřumbe]: «déroulement»
/r/>[ɹ]	Paraguay, Chili	[ɹio]: «fleuve»
/r/final disparait	Venezuela	[amo]: «amour»
/r/et/r̄/ forts	Castille, Pérou	[amor],[deřumbe]

/r̄/ > [l]	Venezuela	[koler]: «courir»

- différents traitements de l'occlusive sonore /d/

indice	Pays/région	exemples
/d/ final>[θ]	Castille	[kantiδaθ]: «quantité»
/d/final disparait	Venezuela, Andalousie	[kantiδa]
/ado/final>[ao]	Chili (Nord), Andalousie	[alusinao]: «alucinés»
/ado/ >[aho]	Colombie	[alusinaho]

- traitement du groupe consonantique complexe /kst/

indice	Pays/région	exemples
/kst/>[kst]	Mexique	[ekstraŋamente]: «étrangement»
/kst/> [st]	Andalousie	[estraŋamente]

- traitement des voyelles devant consonne nasale

indice	pays	exemples
nasalisation	Colombie, Mexique	[semẽhãnte]: « pareil » [semexãnte]

- traitement des voyelles

indice	pays	exemples
Ouverture des voyelles	Venezuela, Chili (centre)	[e̯nfe̯rma]: «malade»

4.3.2. Indices prosodiques

Les sujets ont unanimement désigné l'intonation comme l'indice le plus important pour l'identification des différents parlers[Qui 88].

- différentes configurations finales de phrases déclaratives.

descendante	Andalousie, Mexique, Colombie, Paraguay
modulée	Argentine, Chili (centre et Nord), Pérou
suspensive	Venezuela,

- différentes marges de variations tonales

importante	Paraguay, Argentine, Chili
moyenne	Venezuela, Mexique, Colombie
restreinte	Pérou, Andalousie, Castille

- traitement de l'accent tonique

fort	Paraguay, Argentine, Chili
faible	Castille

- traitement de la durée

Rythme rapide	Andalou Castille, Venezuela
Rythme moyen	Mexique, Argentine, Chili (Nord)
Rythme lent	Pérou, Colombie, Chili (centre) Paraguay
Allongement de la syllabe tonique	Paraguay, Colombie
Rythme syllabique	Mexique

5. Conclusion

Cette étude sur l'identification de différentes variétés d'Espagnol à partir de tests perceptifs a clarifié les points suivants:les meilleurs scores sont obtenus pour l'identification des variétés proches de celle du sujet. Toutefois les frontières entre certaines variétés sont difficiles à établir et certains indices indiqués par les sujets sont valables pour plus d'un pays. Les indices suprasegmentaux sont de loin les plus pertinents selon les sujets.

Au cours de notre recherche nous tenterons d'évaluer le potentiel identificateur de chaque indice et d'établir l'importance de chacun pour l'Identification automatique des parlers espagnols.

Ce travail a été fait dans le cadre d'une convention ARC n°98-02,n°226 du ministère de l'éducation nationale.

Bibliographie

[Bar 99] Barkat, M.,(1999),«Identification of arabic dialects and experimental determination of acoustic cues», ICPhS, San Francisco, Vol 2, pp.901-904.

[Lap 85] Lapesa, R.,(1985),«Origenes y expansion del español atlantico», Rabida,2.

[Lip 94] Lipski ,J.M., (1994) ,«Latin american spanish»,Longman, London and New-York.

[Qui 88] Quilis A.,(1988),«Fonética acústica de la lengua española»ed. Gredos, Madrid.

Perception de la voix parlée : Timbre local et timbre global

Blas Payri

Groupe Traitement du Langage Parlé
LIMSI –CNRS, BP 133 91403 Orsay, France
Tél.: ++33 (0)169 85 80 67 - Fax: ++33 (0)169 85 80 88
Mél: blas@limsi.fr - http://www.limsi.fr/Individu/blas

ABSTRACT

This contribution aims at defining the notions of global timbre and local timbre for speech. From a sentence we chop off two syllables. Listeners must classify 20 sentences and 30 syllables. The results show that syllables are not classified with the same criteria than sentences, and that the identity of the speaker is less relevant than the stress associated with the syllable. In another experiment, listeners must rate the sentences and syllables with predefined criteria (gender, age, tension...) : the results show that there may be a significant difference between the perception of the whole sentence and the ratings for the extracted syllables. There is a global timbre (sentence) made of a succession of local timbres (syllables) that may be very different.

1 INTRODUCTION

La notion de timbre, surtout quand il s'agit de la voix parlée, est une notion dont la définition est peu stable. Castellengo [Cas94] remarque que le mot timbre peut revêtir des sens très différents *« nous parlons du timbre de la voix d'une personne en le comparant à celui d'une autre personne, ou encore des différents timbres d'une même personne, ou encore du timbre de la voix parlée par comparaison à celui de la voix chantée. Il est question aussi du timbre de chaque voyelle, quel qu'en soit le locuteur »* Le timbre concerne donc : 1 l'ensemble des caractéristiques acoustiques qui permettent d'identifier la personne (individualité vocale), 2 celles qui nous permettent de qualifier les sons, principalement dans le domaine spectral, lorsqu'on parle de voix claire, sourde, perçante, nasale, etc. (Castellengo emploie le terme de « sonorité »). Deux approches sont couramment employées dans la littérature : une définition du timbre est ce qui définit un locuteur, et donc le timbre est intimement lié à ce qu'on pourrait appeler l'individualité de la voix. Une deuxième définition concerne les caractéristiques d'une émission vocale donnée : le plus souvent il s'agit de voyelles tenues. Cette dernière définition admet qu'il peut y avoir plusieurs timbres pour un locuteur donné, dépendant du mode de phonation, l'effort vocal, et même les phonèmes employés pour émettre un son, aussi bien que des caractéristiques propres au locuteur – taille des résonateurs, pathologies vocales.

Dans la littérature, on trouve rarement une discussion explicite de la définition employée pour le timbre : souvent les expériences cherchent à caractériser les « sonorités » des émissions vocales, mais en ne prenant qu'un échantillon par locuteur. Implicitement, on cherche donc l'individualité vocale, en faisant l'hypothèse que la sonorité d'un échantillon vocal suffit à étudier l'individualité vocale.

La littérature concernant la synthèse de parole aborde souvent le timbre du point de vue des variations intra-locuteur. En partant des caractéristiques d'un locuteur type, les expériences cherchent à modéliser les variations de qualité de voix à l'intérieur d'une phrase - souffle, craquement, voisement modal (travaux de Klatt [Kla90] et Fant [Fan91]). Par exemple, Sluijter et van Heuven [Slu97] ont mis en évidence que les syllabes toniques du néerlandais et de l'anglais sont caractérisées avant tout par un changement de la pente spectrale et du quotient d'ouverture.

Nous allons définir dans cet article le timbre comme l'ensemble des caractéristiques perceptives associées à un échantillon donné. Ces caractéristiques peuvent concerner le locuteur (âge, sexe), des qualités abstraites (souffle, craquement), des caractéristiques psychologiques, hédonistes... Nous allons essayer de montrer que ces caractéristiques ne sont pas les mêmes si on prend un échantillon long ou une partie, et que les différentes parties d'un échantillon peuvent avoir des caractéristiques différentes, créant une opposition entre timbre global et timbres locaux.

2 MÉTHODES DE RECHERCHE

Pour dégager l'ensemble des caractéristiques perceptives d'échantillons de voix on peut utiliser l'écoute holistique ou l'écoute par axes.

➢ L'écoute holistique : Dans les expériences d'écoute holistique, on demande aux sujets de juger la similarité globale entre les échantillons qui leur sont proposés. Le but de ces expériences est de disposer d'une description de l'espace perceptif sans a priori (Walden [Wal78]), avec l'obtention de paramètres et de stratégies perceptives « proches de l'écoute quotidienne de tout un chacun » ([Wal78]). L'écoute holistique est donc utile pour connaître les critères perceptifs les plus saillants pour un ensemble de sons, mais on a également démontré qu'elle était très dépendante de l'ensemble d'échantillons utilisé.

➢ L'écoute analytique ou par axes : Ce genre d'expériences part de caractéristiques perceptives ou d'axes acoustiques, dont les auteurs cherchent à mieux comprendre les indices, la validité, et l'influence croisée des différents axes. La recherche dans la pathologie vocale utilise énormément cette approche : le but est de créer un outil perceptif d'évaluation des pathologies vocales qui soit objectif et complet valable pour différents groupes d'auditeurs experts, complet dans sa description des changements sonores entraînés par chaque pathologie. Dans les expériences, les auditeurs doivent juger les échantillons selon chaque axe proposé ; on peut ainsi dégager les axes qui servent le mieux à discriminer les échantillons (Kreiman et Gerratt [Kre96]). On obtient ainsi, après analyse des résultats, un ensemble de descripteurs minimal. D'autres expériences se concentrent sur un critère donné, comme l'âge ou le sexe (voir [Pay00] pour discussion).

3 LE MATÉRIAU SONORE

Nous avons choisi d'utiliser la base EUROM pour l'italien. Les enregistrements de cette base ont pour avantage d'avoir été faits dans des conditions identiques pour tous les locuteurs, suivant un protocole strict. Par ailleurs tous les locuteurs adoptaient un niveau neutre de force de voix, avec un ton de lecture neutre pour tous les locuteurs, et un voisement normal. Ce type de bases permet de se focaliser sur les différences entre locuteurs.

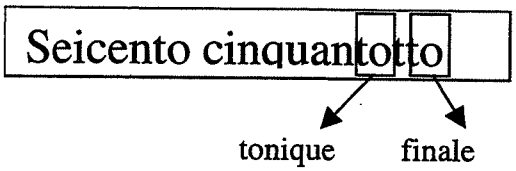


Figure 1 Le matériau sonore était composé de la phrase italienne "seicento cinquantotto" (signifiant "six cents cinquante-huit"). De cette phrase, les deux dernières syllabes "to" étaient extraites : on obtenait une syllabe en position tonique et une syllabe en position finale.

Pour chaque phrase, nous avons extrait deux syllabes, comme illustré dans la figure 1. Nous disposons ainsi, pour chaque locuteur, d'un matériau long (phrase) qui permettra une perception globale, et deux échantillons courts (syllabes) qui obligeront l'auditeur à une perception locale. On peut justifier cette opposition entre local et global comme suit :

- La syllabe est une unité insécable de parole : on peut artificiellement segmenter une syllabe en phonèmes ou diphones, mais un locuteur humain produira toujours une syllabe au minimum (la syllabe la plus simple étant composée d'une voyelle). Certaines expériences utilisent un matériau de type syllabe : par exemple [Wal78] utilisent le mot monosyllabique « beans » dans une expérience de reconnaissance. Beaucoup d'expériences de perception utilisent des voyelles tenues, ce que nous avons écarté car nous voulions garder des extraits de parole continue. Toutes les syllabes ne sont pas isolables dans le contexte d'une phrase à cause des phénomènes de coarticulation, surtout quand il y a continuité de voisement entre les différentes syllabes ;

- On peut considérer que la phrase que nous avons prise est d'une longueur suffisante pour que l'auditeur puisse se faire une image globale du locuteur. Par exemple, Schmidt-Nielsen et Stern [Sch85] montrent que la reconnaissance du locuteur s'améliore asymptotiquement avec la durée de l'échantillon, pour atteindre un maximum vers 2 ou 3 secondes, ce qui est la durée d'une phrase courte.

20 locuteurs ont été choisis dans la base. Parmi les 20 phrases « seicento cinquantotto », 30 syllabes ont été choisies. Une expérience préliminaire a montré que l'ensemble des 20 syllabes toniques et 20 syllabes finales était trop volumineux pour pouvoir être traité dans une expérience de classification libre, en effet les auditeurs doivent tenir compte du matériau dans son ensemble pour décider des classes à faire. Nous avons éliminé les syllabes qui présentaient le plus de variation, notamment les syllabes finales dévoisées, pour retenir 18 syllabes toniques et 12 finales.

Le matériau que nous utilisons rencontre des restrictions : il est utile pour notre expérience d'étude préliminaire des niveaux de perception du timbre, mais par contre, il est insuffisant si nous visons une étude exhaustive du timbre de la voix. Nous disposons en effet de peu d'éléments : un ensemble de 20 locuteurs différents est forcément peu représentatif de l 'espace du timbre. Par ailleurs, il y a peu de diversité intra-locuteur : il s'agit d'une lecture neutre de chiffres, ce qui restreindra le nombre de dimensions. Nous étudions les syllabes finales, et nous pouvons supposer qu'il y a une relation entre la perception globale et la position des syllabes dans la phrase : on pourrait obtenir des résultats différents avec des syllabes en début de phrase. Il faut aussi noter que nous avons des auditeurs francophones sur un matériau italien qui peuvent percevoir les qualités subjectives de façon différente des auditeurs italiens. Cependant certaines recherches montrent que des facteurs perceptifs peuvent ne pas dépendre de la langue du locuteur et de l'auditeur : par exemple Braun et Cerrato [Bra99] ont comparé les réponses d'auditeurs allemands et italiens sur des phrases en allemand et en italien, pour montrer qu'il n'y avait pas de différences significatives dans les estimations d'âge ni entre les groupes.

4 EXPÉRIENCE D'ÉCOUTE HOLISTIQUE

Cette expérience a pour but de dégager les axes perceptifs principaux de notre ensemble d'échantillons, c'est pourquoi nous demandons aux auditeurs de juger la similarité globale entre échantillons sans leur donner de directives de stratégie ou de critères à privilégier.

Tâche

Les auditeurs devaient réaliser une classification libre des échantillons, à l'aide d'un logiciel créé pour l'expérience. D'abord les auditeurs classaient les syllabes « to » puis, dans une étape indépendante, les phrases. Une fois la classification réalisée, le logiciel demandait la stratégie globale suivie, puis, pour chaque classe réalisée, il était demandé une liste de qualificatifs, indiquant ce qui distinguait les sons de cette classe par rapport aux autres. Cette tâche de verbalisation libre permet une compréhension des stratégies des auditeurs.

4.1 Distance holistique

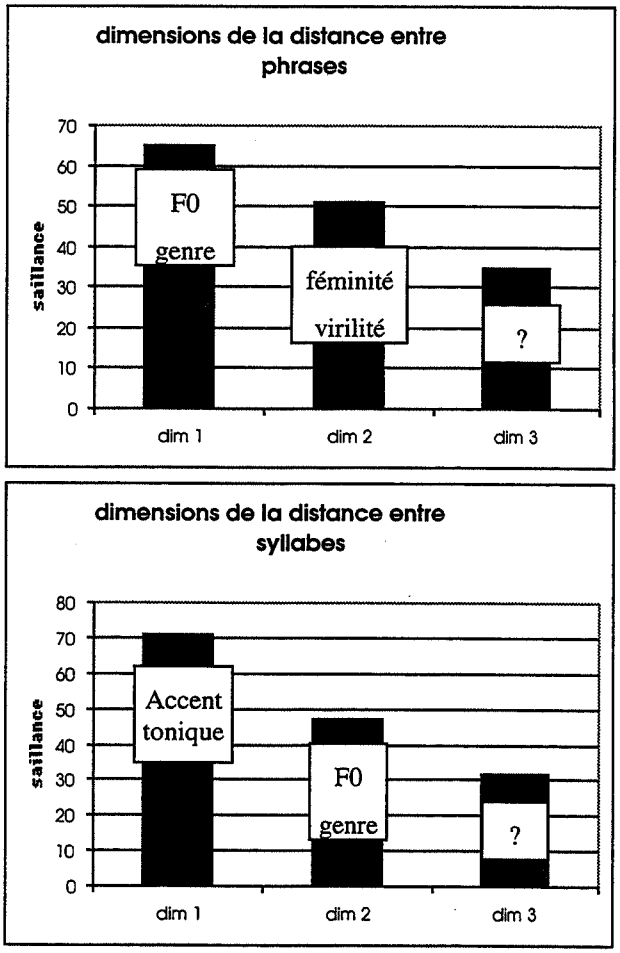


Figure 2 : Les dimensions provenant de l'analyse multidimensionnelle (INDSCAL) pour la distance holistique entre syllabes et entre les phrases. Nous avons indiqué pour chaque dimension les axes avec la plus forte corrélation : ces axes peuvent concerner une mesure acoustique (F0), une caractéristique connue du locuteur (sexe), ou des évaluations perceptives obtenues par des expériences d'écoute par axes (virilité, sexe).

À partir des classifications faites par les auditeurs, nous avons créé une distance holistique (ou de similarité globale), en comptant, pour chaque couple de sons, le nombre de fois où ils apparaissent dans des classes différentes pour chaque classification ; on obtient une matrice vérifiant toutes les propriétés d'une distance. Pour comprendre les éléments de cette distance, nous avons réalisé une analyse multidimensionnelle sur les distances entre phrases et entre syllabes (Figure 2). Pour expliquer les dimensions obtenues, nous avons fait des corrélations avec des mesures acoustiques (F0, jitter, pente spectrale, puissance, rapport énergie harmonique/bruit...) ainsi qu'avec les réponses données par d'autres auditeurs sur le même ensemble de sons dans l'expérience décrite dans le paragraphe 5. Nous constatons que la hauteur (et le sexe qui lui est lié) est un critère très saillant, ce qui se retrouve dans les différentes expériences décrites dans la littérature ([Mur78, 80], [Kre96]). Mais nous constatons que, pour les syllabes, le critère de la position de la syllabe dans la phrase (position tonique ou finale) était plus saillant que la hauteur. En fait, on peut constater que les syllabes étaient classées principalement sur leur accent tonique, ce qui fait que les syllabes provenant du même locuteur (et donc avec des accents toniques différents) étaient classées à part : la position dans la phrase est plus saillante perceptivement que l'individualité vocale du locuteur.

4.2 Résultats de la verbalisation

Nous avons classé les qualificatifs que donnaient les auditeurs pour décrire les classes qu'ils avaient faites en plusieurs catégories (sexe, hauteur, qualité de voix, prosodie, autres). Nous remarquons dans la figure 4 que les auditeurs n'ont pas utilisé les mêmes critères : ils ont privilégié le critère du sexe pour la phrase, alors que la hauteur était plus saillante pour la syllabe, en effet, on peut supposer que le matériau long permet de se faire une image du locuteur, et donc du sexe. La prosodie était plus utilisée pour la phrase, car la syllabe est un support trop court pour des motifs prosodiques.

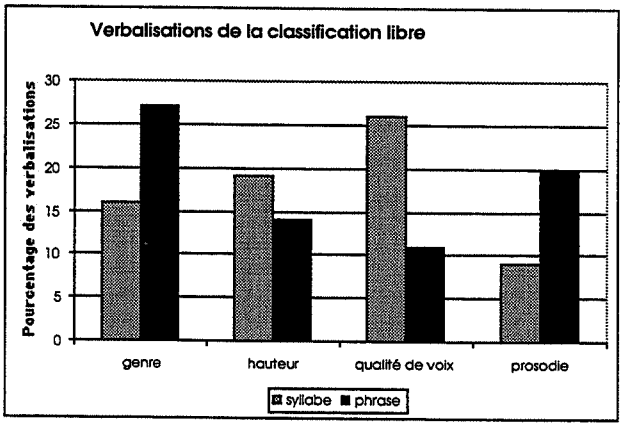


Figure 4 : Les auditeurs devaient donner des qualificatifs servant à décrire les classes qu'ils avaient faites. Nous avons classé ces qualificatifs en types les plus fréquents (sexe, hauteur, qualité, prosodie) et nous avons calculé les fréquences de chaque type.

5 EXPÉRIENCE D'ÉCOUTE PAR AXES PRÉDÉFINIS

Pour mieux comprendre les différences de perception entre le matériau long et le matériau court, nous avons réalisé une expérience supplémentaire, où les auditeurs

devaient juger le même matériau que précédemment (syllabes « to » et phrases) mais cette fois-ci en utilisant des critères prédéfinis, évalués sur des échelles graduées de 1 à 7. Nous avons fait les moyennes des réponses données, et nous constatons que des différences sensibles peuvent apparaître entre les estimations faites pour les phrases et pour les différentes syllabes. Les critères utilisés provenaient des verbalisations libres de l'expérience d'écoute holistique, que nous avons complétés par des critères couramment utilisés dans la littérature. Ces critères étaient : agréable-désagréable, homme-femme, grave-aigu (pour un homme ou une femme), viril-peu viril, féminin-peu féminin, âge minimum , âge maximum, rapide-lent, avec énergie-sans énergie, nasal-non nasal, avec souffle-sans souffle, voilé-clair, tendu-détendu, puissant-faible, bonne prononciation-mauvaise prononciation, vulgaire-distingué, sans prétention-prétentieux, sympathique-antipathique.

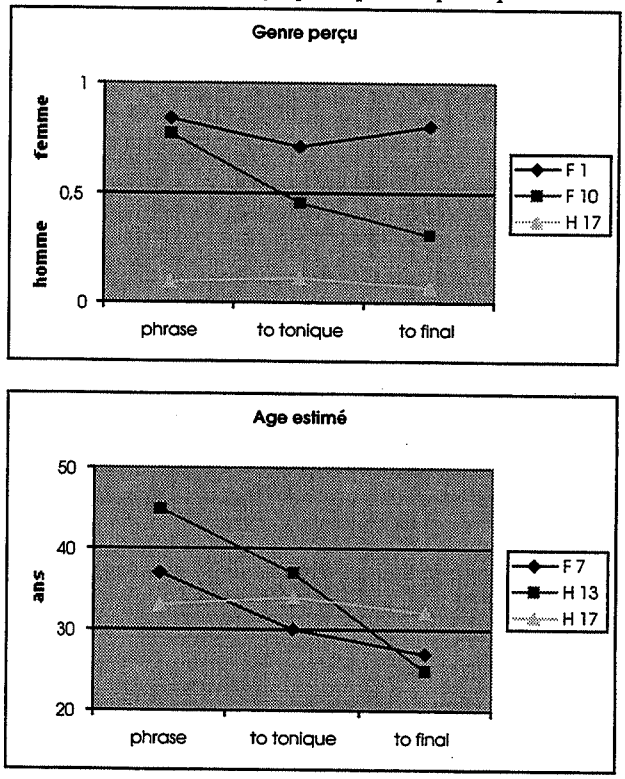


Figure 5 : Moyenne des évaluations des auditeurs pour une phrase d'un locuteur donné, et pour les syllabes tonique et finale de la même phrase. Les valeurs ont été normalisées de 0 à 1 à partir des échelles de 1 à 7 (sauf pour l'âge).

Par exemple, il est bien connu que la prosodie entraîne des variations dans la hauteur ; cependant, nous remarquons dans la figure 5 que quand les auditeurs jugent le sexe, ils peuvent également donner des estimations opposées pour la phrase en entier, ou pour une syllabe issue de cette phrase, comme c'est le cas pour la locutrice F10 (la phrase est perçue comme venant d'une femme, mais les syllabes comme venant d'un homme). Nous trouvons le même phénomène pour la perception de l'âge : par exemple le locuteur H17 a un âge estimé dans la trentaine pour la phrase et les syllabes, mais le locuteur H13 a un âge estimé dans la quarantaine pour la phrase, et dans la vingtaine pour la syllabe finale. On peut expliquer ces phénomènes par les variations de qualité de voix (tension et souffle) au cours d'une phrase.

6 Conclusion

Les expériences que nous avons décrites mettent en évidence que les écoutes sont différentes pour un matériau long (timbre global) et un matériau bref (timbre local) :

- À l'intérieur des caractéristiques globales (phrase) nous pouvons avoir de larges variations des caractéristiques locales (syllabes), comme le montrent les différences de perception de sexe, d'âge, de hauteur, de tension... entre les estimations des phrases et des syllabes extraites. Cependant nous n'avons pas obtenu un modèle prédisant la perception globale à partir des perceptions locales ;

- Les variations de timbre local à l'intérieur d'une phrase sont perceptivement plus saillantes que les caractéristiques du locuteur, car les syllabes du même locuteur sont classées à part si elles n'ont pas la même position dans la phrase.

- Il est donc important, quand on fait de la synthèse de parole ou de la perception des qualités de voix, de tenir compte des variations de timbre autour de valeurs moyennes : une grande partie de la perception des caractéristiques globales peut provenir des variations des valeurs locales tout autant que de leurs valeurs moyennes.

Bibliographie

[Bra99] Braun, A. & Cerrato, L. (1999), " Estimating speaker age accros languages ", XIV ICPhS, pp 1369-1372.

[Fan91] Fant, G., Kruckenberg, A. & Nord, L. (1991), " Prosodic and segmental speaker variations ", Speech Communication, 10, pp 521-531.

[Kla90] Klatt, D.H. & Klatt, L.C. (1990), " Analysis, synthesis, and perception of voice quality variations among female and male talkers ", JASA, 87, (2).

[Kre96] Kreiman, J. & Gerratt, B.R. (1996), " The perceptual structure of pathologic voice quality ", JASA, 100, (3).

[Pay00] Payri, B. (2000) "Perception de la voix parlée: la cohérence du timbre du locuteur", Mémoire de thèse, Université ParisXI.

[Slu96] Sluijter, A.M. & van Heuven, V.J. (1996), " Spectral balance as an acoustic correlate of linguistic stress ", JASA, 100, pp. 2471-2485.

[Slu97] Sluijter, A.M., van Heuven, V.J. & Pacilly, J.J.A. (1996), " Spectral balance as a cue in the perception of linguistic stress ", JASA, 101 (1).

[Sch85] Schmidt-Nielsen, A. & Stern, K.R. (1985), " Identification of known voices as a function of familiarity and narow-band coding ", JASA, 77 (2).

[Wal78] Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A. & Schwartz, D.M. (1978), " Correlates of psychological dimensions in talker similarity", JSHR, 21, pp. 265-275.

Perception de la voix parlée : La cohérence des caractéristiques vocales du locuteur

Blas Payri

Groupe Traitement du Langage Parlé
LIMSI–CNRS, BP 133 91403 Orsay, France
Tél.: ++33 (0)169 85 80 67 - Fax : ++33 (0)169 85 80 88
Mél : blas@limsi.fr - http://www.limsi.fr/Individu/blas

ABSTRACT

We present an experiment of substitution of syllables from different speakers within a sentence. The results show that 23% of the mixes are accepted, and up to 51% of the mixes with same gender speakers. The main criterion is pitch, specially the respect of prosodic patterns. A second experiment with pitch transformation confirms the importance of pitch, but shows that voice quality must not differ substantially for the mixture to be accepted.

1 INTRODUCTION

La notion de timbre, que ce soit dans le domaine de la voix ou de la perception des instruments, est généralement définie comme l'ensemble des caractéristiques sonores qui permettent de distinguer soit deux échantillons, soit deux locuteurs (ou deux instruments dans le domaine instrumental). Beaucoup d'expériences cherchent à connaître les caractéristiques permettant de distinguer les locuteurs [Mat73], à partir d'un ensemble d'échantillons de voix pour chaque locuteur. Pourtant, Kreiman et al. [Kre91] montrent qu'il y a des confusions quant au locuteur pour plus de 20% des échantillons, même quand il s'agit de phrases longues. On peut donc s'interroger sur l'existence d'un "timbre individuel" qui permettrait de distinguer les locuteurs.

Dans cet article, nous ne posons pas a priori que les échantillons de parole sont différenciables par rapport au locuteur, mais nous cherchons à savoir quelles caractéristiques perceptives font qu'un ensemble d'échantillons concaténés proviennent du même locuteur. Nous introduisons la notion de cohérence des caractéristiques vocales : cette notion cherche à définir les limites que doivent respecter toutes les composantes d'une élocution, pour qu'elle semble avoir été prononcée par une seule personne réelle.

2 RESSEMBLANCE ET SUBSTITUABILITÉ

Dans une expérience précédente [Pay00], nous avons utilisé comme matériau la phrase « seicento cinquantotto ». De cette phrase, nous avions extrait les deux dernières syllabes « to », puis nous avons demandé aux auditeurs de classer les syllabes et les phrases selon leur similarité globale. Un des résultats de cette expérience a été que les syllabes les plus semblables n'étaient pas celles qui provenaient du même locuteur (on disposait d'une syllabe « to » tonique et une syllabe « to » finale pour chaque locuteur), mais plutôt, le critère de ressemblance était la position dans la phrase : les variations d'effort vocal dues à la prosodie entraînaient des variations de timbre plus importantes que les différences interlocuteur.

La question que nous avons cherché à résoudre était de savoir si cette ressemblance entre syllabes était en fait une équivalence, en d'autres termes, si on pouvait substituer une syllabe par une autre lui ressemblant. Nous disons que deux syllabes sont semblables si elles sont catégorisées ensemble ; et que deux syllabes sont substituables si leur permutation dans leurs phrases d'origine résulte en de nouvelles phrases acceptables, i.e. que la syllabe introduite respecte la cohérence des caractéristiques vocales du locuteur.

3 EXPÉRIENCE DE MONTAGE

3.1 *Contraintes pour la substitution*

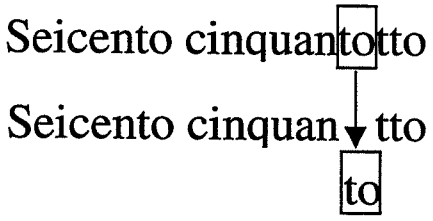


Figure 1 Le montage est fait en sélectionnant la syllabe « to » tonique dans une phrase, et en l'introduisant dans une autre phrase en remplacement de la syllabe tonique équivalente. Le montage est fait sans transformation supplémentaire du son.

Dans notre expérience de substitution de syllabes, nous cherchons à déterminer les paramètres de la cohérence des caractéristiques vocales du locuteur. Cependant, il y a d'autres contraintes de source, comme le montrent les expériences dans le domaine de l'analyse de scènes auditives (Bregman, [Bre90]).

- Contraintes d'environnement : si les segments que l'on concatène ont été enregistrés dans des conditions différentes (bruit ambiant, distance au microphone, système d'enregistrement), on percevra immédiatement qu'il y a un montage, car il est

impossible qu'une personne change d'environnement de façon abrupte. Le matériau que nous avons choisi provient de la base EUROM, qui a été enregistrée avec un protocole strict quant à l'environnement d'enregistrement ; on peut donc considérer que nous sommes affranchis des contraintes d'environnement.

- Contraintes de voisement : on ne doit pas avoir de changements abrupts dans les courbes de F0, ni dans les courbes de formants. Pour éviter ce genre de discontinuités, nous avons choisi de prendre une syllabe « bien détachée », encadrée par deux consonnes plosives sourdes : il s'agit de la syllabe « to » tonique dans la phrase « seicento cinquantotto », comme illustré dans la figure 1. Ce type d'expérience est donc difficilement réalisable avec d'autres types de matériau qui ne permettent pas une segmentation nette.
- Contraintes de la langue : ceci comprend le contenu linguistique (on ne peut pas introduire un segment qui est contradictoire avec le mot attendu), le contenu prosodique (par exemple, si on prend deux syllabes « to » d'une même phrase, et qu'on les permute, il y a rejet, car la prosodie est brisée), et la coarticulation. C'est pourquoi nous avons choisi de ne remplacer des syllabes que par d'autres syllabes ayant la même position dans une phrase équivalente (voir figure 1).

Nous voyons donc que la cohérence des caractéristiques vocales du locuteur n'est qu'une contrainte qui s'ajoute aux autres contraintes : si le montage est accepté, on peut dire qu'il y a cohérence vocale, mais dans le cas contraire, il se peut qu'une autre des contraintes du montage soit violée.

3.2 *Conditions de l'expérience*

Nous avons utilisé 20 locuteurs de la base EUROM pour l'italien, ce sont les mêmes locuteurs que nous avions étudiés dans une expérience préalable de classification libre et d'écoute par axes. Nous avons fait toutes les combinaisons des syllabes « to » toniques et des phrases, ce qui résulte en 380 montages et 20 phrases non montées. Les sons ont été présentés sur un questionnaire internet par pages indépendantes de 40 sons, avec pour chaque page 20 phrases montées et les 20 phrases non montées, ce qui représentait 760 écoutes. Nous avons retenu 22 auditeurs, dont 13 ont écouté les sons avec des haut-parleurs dans une cabine isolée, et 9 ont utilisé des casques dans un bureau silencieux.

3.3 *Tâche*

Pour chaque élocution, les auditeurs indiquaient s'ils entendaient un locuteur (acceptation) ou un montage (rejet). On a obtenu 22 réponses par montage. À partir des réponses, on calculait un indice de distance, que nous appellerons distance de montage, en calculant le nombre de rejets sur le nombre total de réponses obtenues : une distance proche de 1 indique un montage majoritairement rejeté, et pour les montages acceptés la distance est proche de 0.

4 MONTAGE : RÉSULTATS

Nous avons posé comme critère d'acceptation d'un montage que la distance de montage soit inférieure à 0,5 (c'est-à-dire que la majorité des auditeurs l'ont accepté comme étant une phrase non montée). Nous pouvons voir dans la figure 2 que 23 % des montages sont acceptés, et que ce pourcentage augmente si on se restreint aux montages faits entre locuteurs de même sexe : la moitié des montages entre voix de femmes est acceptée.

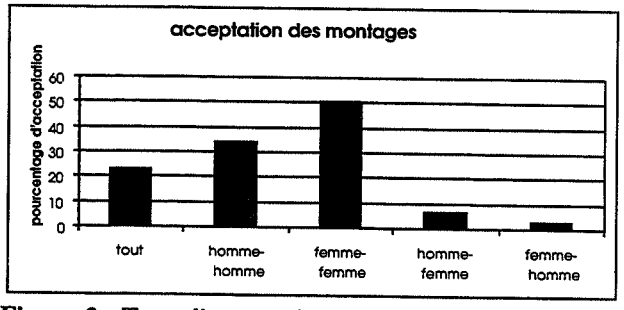


Figure 2 : Taux d'acceptation des montages, en fonction du sexe du locuteur de la phrase porteuse et de la syllabe remplaçante.

L'expérience de montage montre qu'une syllabe, provenant d'un locuteur donné, peut en fait être insérée dans des phrases provenant de locuteurs différents, résultant en une nouvelle phrase montée acceptable. Ce résultat montre qu'il n'y a pas de timbre individuel du locuteur au niveau de la syllabe. Pour mieux comprendre les conditions d'acceptation nous allons procéder à une analyse de la distance perceptive de montage.

4.1 *Dimensions de la distance*

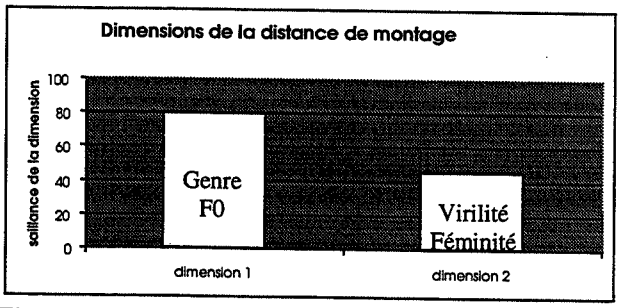


Figure 3 : Saillance des dimensions obtenues par l'analyse multidimensionnelle (INDSCAL) de la distance de montage entre tous les locuteurs. Deux dimensions suffisent pour expliquer 81 % de la variance totale de la distance de montage.

L'indice de distance obtenu n'est ni symétrique, ni réflexif : nous avons donc symétrisé les valeurs de la matrice (par simple moyenne) et annulé sa diagonale afin d'opérer une analyse INDSCAL. Pour expliquer les dimensions, nous avons calculé les corrélations avec différents axes acoustiques (F0, pente spectrale, jitter...) et avec des jugements perceptifs obtenus dans une

expérience précédente concernant les mêmes sons (virilité, agrément, raucité, âge perçu...). Nous pouvons observer dans la figure 3, que la première dimension obtenue, qui est de loin la plus saillante, est liée au sexe et à F0, la deuxième dimension est également liée à F0, mais de façon inverse pour les hommes (virilité) et pour les femmes (féminité). Pour éliminer le critère évident du sexe, nous avons ensuite analysé les distances de montage en ne considérant que les montages faits avec des locuteurs de même sexe, comme illustré dans la figure 4.

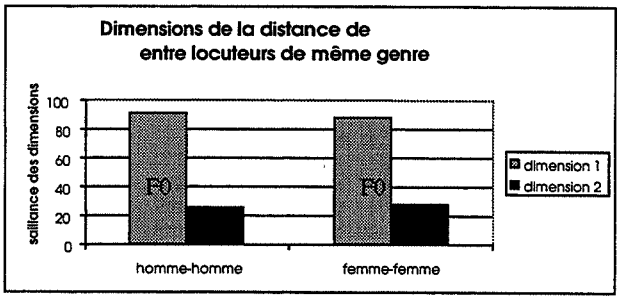


Figure 4 : Dimensions obtenues après analyse multidimensionnelle de la distance entre locuteurs de même sexe. Dans les deux cas, deux dimensions suffisaient à expliquer 85 % de la variance de la distance pour les montages de locuteurs de même sexe.

Nous voyons que même en éliminant le sexe, la première dimension, de loin la plus importante, demeure la hauteur : nous voyons dans la figure 5 que la hauteur à elle seule explique la plupart de la variance de la distance de montage (84 % de la variance pour les hommes et 77 % pour les femmes).

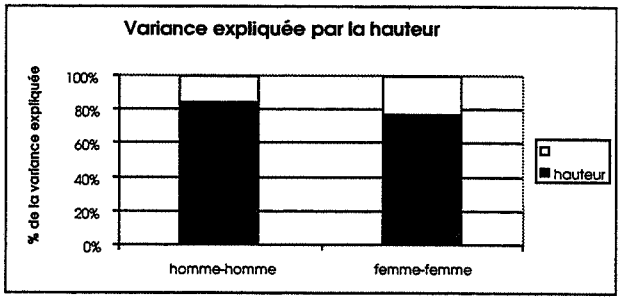


Figure 5 : Variance de la distance de montage entre locuteurs de même sexe, expliquée par le paramètre F0

Nous observons donc que la qualité de voix revêt peu d'importance, et qu'au contraire la première dimension est la hauteur. On peut rapprocher ces résultats des recherches faites en reconnaissance et discrimination du locuteur : par exemple Kreiman et Precoda [Kre91] montrent que la hauteur moyenne est le premier facteur dans l'individualité du locuteur, ce qui est corroboré par la littérature dans le domaine.

4.2 Prosodie et distance de montage

On peut distinguer deux effets dans le rejet dû à la hauteur :

➢ Si la syllabe introduite a une hauteur très différente de la hauteur moyenne d'un locuteur, on pourra penser qu'il y a rupture de la cohérence du locuteur (cette syllabe est impossible pour le locuteur).

➢ Si par contre la syllabe introduite reste dans l'ambitus habituel du locuteur, on peut penser qu'il y a rupture de la prosodie : c'est-à-dire que le locuteur peut émettre cette syllabe, mais que le contexte de la phrase supposerait une autre hauteur. Pour mieux comprendre les effets de la prosodie nous avons tracé les contours mélodiques des différents montages, comme illustré dans la figure 6.

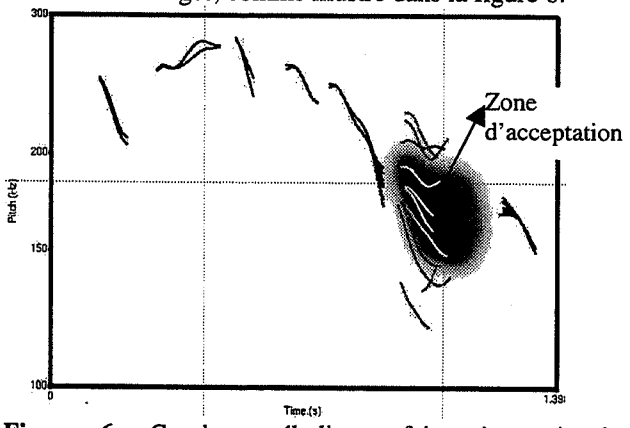


Figure 6 : Courbes mélodiques faites à partir des différents montages sur la phrase 6. On peut définir une zone d'acceptation, dans laquelle toutes les syllabes sont acceptées.

Nous pouvons voir que toutes les syllabes dont les hauteurs sont situées autour d'une certaine valeur sont acceptées, et les autres rejetées : on peut alors définir une zone d'acceptation. Pour qu'une syllabe introduite soit acceptée, il n'est pas nécessaire qu'elle ait une valeur précise de hauteur, on peut plutôt dire que certaines valeurs sont plus probables que d'autres en fonction du contexte prosodique.

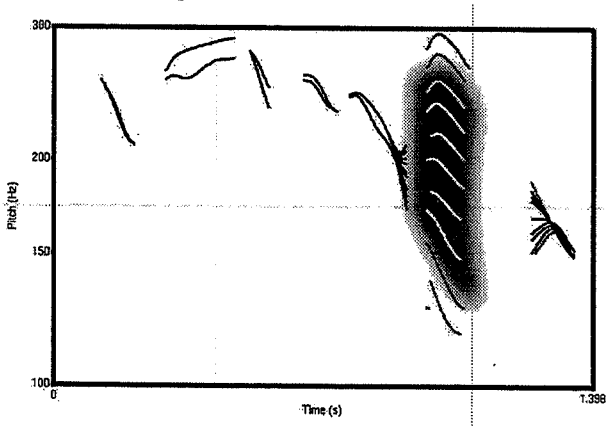


Figure 7 Courbes mélodiques obtenues en transformant progressivement la valeur de F0 de la syllabe originale : on peut créer une zone d'acceptation

5 MONTAGE ET MODIFICATION DE F0

Nous avons établi que les syllabes qui sont dans une « zone d'acceptation » donnent des montages

majoritairement acceptés, et les autres sont rejetées. Nous voulons tester si cette condition est suffisante, i.e. si on ramène une syllabe dans l'intervalle d'acceptation va-t-elle être acceptée ? et question inverse, si nous extrayons une syllabe acceptée de l'intervalle, va-t-elle être rejetée ? Pour cela nous avons fait une expérience complémentaire de montage, avec transformation de F0 avec la méthode PSOLA. Le matériau de cette expérience était :

- Une phrase porteuse (celle de la locutrice 6) avec les syllabes allogènes non modifiées, comme précédemment (19 montages et l'original)
- La phrase porteuse et syllabes allogènes modifiées de façon à avoir le même F0 que la syllabe remplacée (19 montages)
- La phrase porteuse avec syllabe originale modifiée en F0 (10 degrés de modification) (voir figure 7)

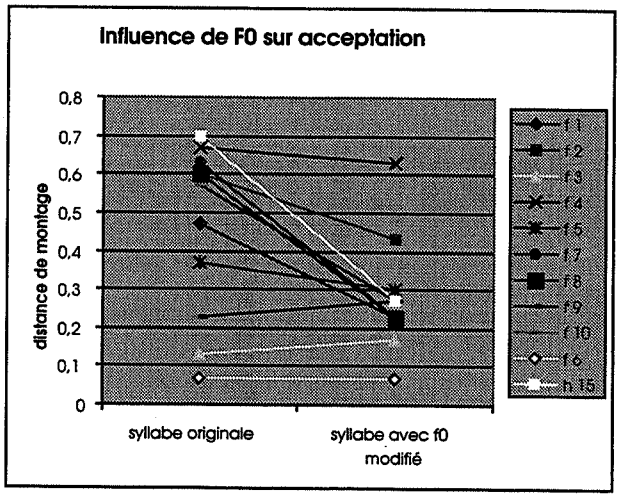


Figure 8 : L'influence de F0 dans l'acceptation des montages : à gauche nous avons la distance de montage pour la syllabe sans transformation, à droite la distance de montage pour la même syllabe transformée en F0 de façon à être dans l'intervalle d'acceptation. On constate que la transformation de F0 améliore globalement l'acceptation.

Dans la figure 7, nous observons que quand une syllabe bien acceptée est modifiée en F0, elle va progressivement être de plus en plus rejeté : on peut alors créer une zone d'acceptation, mais cette zone est plus grande que celle obtenue avec des syllabes provenant de locuteurs différents. On peut donc dire que le fait que la syllabe ait un timbre très compatible (ici c'était la syllabe originale de la phrase) augmente l'ambitus de F0 où elle sera acceptée. Dans la figure 8 nous observons que la transformation de F0 de façon à ce que la syllabe introduite ait la même hauteur que la syllabe remplacée, améliore l'acceptation du montage. Cependant nous avons quelques paradoxes : malgré des hauteurs proches, certains montages sont rejetés. On peut donc dire que le respect de la prosodie est une condition nécessaire mais pas entièrement suffisante. Dans le cas qui nous occupe, les montages rejetés proviennent de locuteurs avec des qualités de voix nettement différentes : la locutrice la plus âgée ne se mélangeait pas aux autres voix, et de même pour une autre locutrice dont la voix était perçue comme très tendue.

6 CONCLUSIONS

Le principal résultat de ces expériences est qu'un timbre local (syllabe) peut être partagé par plusieurs timbres globaux : l'individualité du locuteur ne repose pas sur la description instantanée du signal de parole. Etant donné un contexte, plusieurs syllabes, éventuellement de locuteurs différents sont acceptables. L'aspect dynamique est très important dans l'individualité vocale : nos expériences montrent que la condition première d'acceptation d'un montage est qu'il y ait respect des contraintes de parole, et au premier rang la prosodie. Ces conclusions peuvent rejoindre celles qui sont issues de la reconnaissance du locuteur : Abberton et Fourcin [Abb78] montrent l'importance de la prosodie dans la reconnaissance. Les recherches de Bailly et Morlec [Mor95], [Bai97] montrent qu'on peut modéliser des contours prototypiques pour différentes attitudes valables pour différents locuteurs. L'individualité vocale peut dépendre plus des aspects dynamiques (prosodie) que de valeurs moyennes.

La synthèse par concaténation peut être élargie : si on peut prédire la hauteur et le mode de phonation d'une syllabe dans un contexte, une syllabe d'un locuteur pourra être utilisée pourvu qu'il n'y ait pas de différences notables de mode de phonation, ou de caractéristiques propres comme l'âge ou les pathologies.

BIBLIOGRAPHIE

[Abb78] Abberton, E. et Fourcin, A. (1978) "Intonation and speaker identification", Language and speech, Vol 21, pp. 305-318.

[Bai97] Bailly, G. (1997), "No future for comprehensive models of intonation?", in Computing Prosod : Computational models for processing spontaneous speech, Springer Verlag pp. 157-164.

[Bre90] Bregman, A.S. (1990), "Auditory scene analysis", The MIT Press.

[Kre91] Kreiman, J. et Papcun, G. (1991), "Comparing discrimination and recognition of unfamiliar voices", Speech Communication, Vol 10, pp. 265-275

[Mat73] Matsumoto, H., Hki, S., Sone, T., et Nimura, T. (1973), "Multidimensional representation of personal quality and its acoustical correlates", IEEE transactions Audio Electroacoust., Vol 21, pp. 428-436.

[Mor95] Morlec, Y., Aubergé, V. et Bailly, G. (1995), "Evaluation of automatic generation of prosody with a superposition model", Proceedings of ECSCT, Vol 3, pp. 2043-2046.

[Pay00] Payri, B. (2000), "Perception de la voix parlée : la cohérence du timbre du locuteur", Mémoire de Thèse, Université ParisXI.

Production et pathologies

Un diagnostic phonétique pour les déficiences auditives

A. Bonneau, Parham Mokhtari

LORIA–CNRS & INRIA Lorraine, Bâtiment LORIA, BP 239,
54506 Vandœuvre-lès-Nancy FRANCE.
Tel. (33) 3 83 59 20 80, FAX: (33) 3 83 41 30 79, E-mail:bonneau@loria.f

Abstract

We propose a phonetically-guided diagnosis of auditory deficiency, which hinges on a reasonably small battery of speech-based auditory tests, including the dichotic presentation of split stimuli, and is based on a carefully constructed corpus of synthetic sounds. Our aim is to complement the diagnosis of sensorineural hearing deficiencies, (such as a frequency-dependent rise in the threshold of audibility, a reduced degree of frequency selectivity, and a reduced degree of temporal resolution), in order to improve the correction afforded by auditory prosthesis. To test our method, we simulate a frequency-selective loss of audibility of -40 dB at a given center frequency. The diagnosis will be tested on hearing impaired people at the Central Hospital of Nancy.

1. Introduction

Les déficiences auditives à l'origine des troubles de perception de la parole sont nombreuses et étroitement liées les unes aux autres. Les plus importantes sont l'élévation du seuil d'audibilité, les pertes en sélectivité fréquentielle et en résolution temporelle [5]. Avec le développement de nouvelles technologies, il est maintenant possible d'implémenter dans les aides auditives des techniques récentes de traitement du signal qui améliorent de manière sensible la compréhension de la parole [4].

Notre équipe de recherche, associée à deux équipes médicales spécialisées en audiologie, travaille sur des transformations de signal destinées à apporter des corrections bien adaptées aux déficiences auditives de chaque sujet et respectant, dans la mesure du possible, les oppositions phonétiques entre les sons de parole. Parmi les transformations qui peuvent s'appliquer aux déficits psycho-acoustiques cités plus haut, nous sommes particulièrement intéressés par le renforcement des formants ou d'indices acoustiques importants, ainsi que par une baisse sélective du débit de parole.

Pour que la correction soit précise, il est nécessaire de disposer d'un diagnostic fin. On distingue trois grands types de diagnostic selon qu'ils reposent sur l'audiométrie tonale, l'audiométrie vocale, ou les tests psycho-acoustiques. L'audiométrie tonale recherche les seuils d'audibilité de tons purs dont on fait varier la fréquence. Par conséquent, elle ne fournit aucun renseignement sur les autres types de déficience. Les tests psycho-acoustiques sont capables de fournir des mesures quantitatives de déficiences supraliminaires telles que la perte de sélectivité fréquentielle mais sont souvent longs et pénibles pour les patients. Le diagnostic vocal en revanche nous semble particulièrement intéressant puisqu'il repose sur la tâche plus naturelle de reconnaissance de différentes unités de la parole (sons, mots ou phrases) et s'adapte donc bien à l'objectif principal des aides auditives : améliorer l'intelligibilité de la parole. Notre papier est consacré à la présentation de nos méthodes et de nos outils de diagnostic vocal : un corpus de sons synthétiques, une plate-forme de tests pour un diagnostic vocal, ainsi que la réalisation de tests de simulation de déficiences auditives.

2. L'audiométrie vocale

2.1. Le corpus de sons synthétiques

Nous avons créé un corpus de voyelles synthétiques à deux formants. Les fréquences formantiques de ces voyelles ont été choisies de manière à ce qu'elles couvrent une grande partie de l'espace défini par les fréquences de F1 et de F2 et que les voyelles puissent être regroupées en paires qui ne diffèrent que par la fréquence d'un formant. Cette structure formantique simple facilite l'interprétation des confusions entre les voyelles et, en particulier, permet de mieux cerner les régions fréquentielles dans lesquelles les déficiences auditives du patient sont les plus aiguës.

Les stimuli ont été créés à l'aide de la branche parallèle du synthétiseur de Klatt. Leur durée a été fixée à 200 ms, leur fréquence fondamentale à 120 Hz (voix masculine) et à 220 Hz (voix féminine). Afin d'améliorer le naturel des stimuli, la fréquence fondamentale descend légèrement du début à la fin des voyelles (pente de 5%). De même, l'intensité est légèrement plus faible au début et à la fin de chaque stimulus, ce qui élimine les clicks ou discontinuités brutales dans l'intensité globale du signal. Les fréquences formantiques des voyelles sont présentées dans la table 1 (voix masculine).

2.2. Les tests d'audiométrie

Les tests présentés ci-dessous se dérouleront à l'Hôpital Central de Nancy à l'issue des tests habituels d'audiométrie tonale et vocale. L'équipe d'audiologie les proposera à tous les sujets qui présenteront une au-

diométrie vocale perturbée. Ces sujets devront identifier les voyelles synthétiques de notre corpus sous différentes conditions d'écoute.

- C1. Écoute monaurale sans adjonction de bruit.
- C2. Écoute monaurale avec adjonction de bruit.
- C3. Écoute dichotique sans adjonction de bruit.
- C4. Écoute dichotique avec adjonction de bruit.

En outre, afin d'apprécier la résolution temporelle de chaque sujet, une phrase sera présentée en débit normal et en débit lent (vitesse deux fois moins rapide que la normale). Cette phrase, peu prédictible, sera de surcroît dite en anglais afin de ne pas laisser aux auditeurs d'accès au lexique.

L'écoute dichotique consiste à faire écouter simultanément une partie d'un son à une oreille et l'autre partie à l'autre oreille. Grâce à une expérience d'écoute dichotique où l'un des deux stimuli simultanés est constitué par le fondamental et certains formants d'une voyelle, et l'autre stimulus par le fondamental et les autres formants de cette même voyelle, Carlson *et al.* [2] ont montré que les auditeurs réussissaient à identifier correctement la voyelle décomposée, prouvant ainsi que le timbre vocalique était intégré à un niveau non périphérique. Avec une décomposition du signal sonore judicieuse et un double appareillage, l'écoute dichotique pourrait donc être exploitée pour améliorer la perception des sujets souffrant d'une mauvaise sélectivité fréquentielle. Les expériences de Chaudari *et al.* [3] tendent à renforcer cet espoir. En effet, ces auteurs ont simulé l'effet d'une perte de sélectivité fréquentielle en modifiant le signal de parole et ont montré, avec des sujets normo-entendants, que la présentation dichotique de consonnes améliorait leur reconnaissance, particulièrement en milieu bruité. Pour nos conditions C3 et C4, le premier formant et le fondamental d'une voyelle sont présentés à une oreille, le deuxième formant est présenté à l'autre oreille. Ces tests d'écoute dichotique peuvent nous permettre à la fois de diagnostiquer des problèmes de sélectivité fréquentielle et de vérifier si la présentation dichotique des sons est susceptible d'améliorer l'intelligibilité de la parole chez les sujets malentendants.

Les stimuli sont présentés à différents niveaux de bruit additif. L'adjonction de bruit est nécessaire car les personnes malentendantes, si elles arrivent souvent à comprendre la parole en milieu non bruité, rencontrent de grandes difficultés dans le bruit.

Une plate-forme d'audiométrie, écrite pour Windows et destinée aux audiométristes de l'Hôpital Central, a été réalisée au sein de notre équipe. Les utilisateurs peuvent choisir facilement le type de présentation (monaurale, dichotique), les stimuli, qui apparaissent dans des listes, le niveau de bruit additif ainsi que le niveau d'intensité du stimulus présenté à chaque oreille (qui est fonction du déficit du patient).

2.3. Le diagnostic

Les modifications du débit de parole nous aideront à diagnostiquer des problèmes de résolution temporelle, le ralentissement du débit permettant aux sujets souffrant d'une mauvaise résolution de reconnaître les phrases qu'ils ne peuvent comprendre à vitesse normale. De même, l'écoute dichotique, dans la mesure où elle améliore sensiblement la performance des sujets, peut indiquer des problèmes de sélectivité temporelle. Enfin, ainsi que nous le montrerons dans le chapitre suivant, les pertes d'audibilité locales (limitées à une région fréquentielle donnée) peuvent être appréciées grâce à notre corpus de sons synthétisés. Il faut néanmoins souligner que le diagnostic des déficiences auditives n'est pas une tâche aisée puisque dans la plupart des cas plusieurs types de déficience interviennent en même temps chez un même sujet.

3. LES TESTS DE SIMULATION

Il est possible de simuler l'effet des déficiences psychoacoustiques sur le spectre auditif en modifiant le spectre de parole. Les sujets normo-entendants peuvent ainsi apprécier l'impact des troubles auditifs sur l'intelligibilité des sons de la parole. Bien entendu, les simulations de pertes auditives sont toujours simplificatrices et ne donnent qu'une image imparfaite de ce que perçoivent les personnes malentendantes. Il nous a semblé néanmoins intéressant d'utiliser ces simulations afin d'effectuer un premier test de notre diagnostic dans des conditions simples.

Nous avons donc simulé une perte d'audibilité dans la région du deuxième formant des voyelles. Pour ce faire, nous avons défini une courbe auditive possédant un minimum de -40 dB à une fréquence donnée (1300, 1600, puis 1900 Hz). Des segments linéaires joignent ce point aux points distants de 500 Hz de chaque côté de l'axe fréquentiel et possédant une intensité de 0 dB. Partout ailleurs sur l'axe fréquentiel, les points possèdent également une intensité de 0 dB. Cette courbe auditive est ensuite ajoutée à chaque spectre FFT à court terme des stimuli vocaliques, et le spectre résultant, modifié, est utilisé pour synthétiser le nouveau stimulus par la méthode d'overlap-and-add (OLA).

Ces simulations ont été appliquées aux stimuli caractérisés par un F0 bas (voix masculine, table 1). Nous avons donc généré trois groupes de stimuli, un pour chaque centre fréquentiel. Les trois groupes, ainsi que les stimuli non modifiés, ont été présentés séparément à des sujets normo-entendants.

3.1. Protocole expérimental

Dix auditeurs normo-entendants de langue française ont participé à l'expérience. Ils ont écouté les stimuli dans une pièce calme à l'aide d'un casque Sennheiser HD520 II.

Nous avons demandé aux auditeurs de choisir leur réponse parmi les six voyelles orales /i, E, A, y, OE, O/; /E/ correspondant à /e, ɛ/, /A/ à /a, ɑ/, /OE/ à /ø, œ/ et /O/ à /o,ɔ/. Nous avons donc exclu les distinctions entre les voyelles dont l'opposition est souvent neutralisée en français. Le choix des voyelles du corpus du reste n'impose pas ces distinctions difficiles.

Les auditeurs ont répondu oralement après avoir entendu, à une seconde d'intervalle, deux répétitions de

chaque stimulus; un intervalle de quatre secondes de silence séparant deux stimuli différents. Chacun des dix auditeurs a écouté cinq répétitions de la même voyelle, portant à 50 le nombre de total de réponse par voyelle.

3.2. Résultats

Des expériences de perception réalisées avec des voyelles synthétiques à deux formants [1] ont montré qu'une voyelle d'avant possédant un deuxième formant de faible amplitude était perçue comme une voyelle d'arrière de même degré d'ouverture. Par conséquent, l'observation d'une chute du taux d'identification de certaines voyelles d'avant, confondues avec des voyelles d'arrière de même degré d'ouverture, devrait indiquer une perte de sensibilité (à l'amplitude) dans la région de leur deuxième formant. Ce diagnostic sera d'autant plus sûr que les voyelles de même F1 mais de F2 différents ne seront pas affectées.

La structure très simple de nos stimuli (absence de F3 et de formants supérieurs, qui jouent un rôle dans l'identification des voyelles aiguës, l'organisation des voyelles en paires, ainsi que la répartition régulière des fréquences formantiques) doit faciliter l'interprétation des confusions vocaliques et, en particulier, permettre la localisation fréquentielle de problèmes auditifs. Les résultats de l'expérience tendent à le confirmer.

Les voyelles non modifiées ont été relativement bien identifiées; le score d'identification allant de 75% à 100 % selon les voyelles, à l'exception de la voyelle /ɔ/ (score de 65%) confondue très souvent avec /œ/. Comme le montre la figure 1, les voyelles les plus affectées par notre simulation de perte d'audibilité centrée à 1300 Hz sont les voyelles centrales et antérieures qui possèdent un F2 égal ou légèrement supérieur à cette fréquence. De même, la simulation centrée à 1600 Hz détériore essentiellement l'identification de la voyelle /ø/, dont le F2 est à 1600 Hz, mais aussi celle de /y/ (F2 à 1800 Hz). La simulation centrée à 1800 Hz affecte surtout la voyelle /y/. Conformément à notre attente, les voyelles mal identifiées sont en général confondues avec une voyelle d'arrière, de même degré d'ouverture. Il peut être intéressant de noter que certaines voyelles, (/ɛ/ en particulier) , dont le F2 est affecté par nos simulations, sont toujours relativement bien identifiées. Une interprétation de ce phénomène exigerait d'autres investigations et sort du cadre de cet article. Néanmoins, les résultats globaux de l'expérience permettent bien de localiser la région fréquentielle où le déficit est le plus important.

4. PERSPECTIVES

Nous avons présenté un diagnostic phonétique de déficiences auditives fondé sur des voyelles synthétiques à deux formants et sur une batterie de tests auditifs simples. Nos tests de simulations (très simples) tendent à confirmer l'utilité d'un jeu de stimuli dont les fréquences formantiques sont bien contrôlées. Notre corpus sera complété par l'adjonction de consonnes en contexte vocalique.

Assistés de l'équipe d'audiométrie de l'Hôpital Central de Nancy, nous testerons prochainement l'efficacité de la présentation dichotique et les effets du bruit additif avec les patients de l'hôpital présentant une audiométrie vocale (classique) perturbée.

BIBLIOGRAPHIE

[1] W.A. Ainsworth and J.B. Millar. The effect of relative formant amplitude on the perceived identity of synthetic vowels. *Language and Speech*, 15:328–341, 1972.

[2] R. Carlson, G. Fant, and B. Grandström. Two formant models, pitch and vowel perception. In G. Fant and M.A. Tatham, editors, *Auditory analysis and perception of speech*, pages 55–82. Academic Press, New York and London, 1975.

[3] D.S. Chaudari and P.C. Pandey. Dichotic presentation of speech signal with critical band filtering for improving speech perception. In *Proceedings of ICASSP*, Berlin, 1998.

[4] P.C. Loizou. Mimicking the human ear. *IEEE Signal Processing Magazine*, pages 101–130, 1998.

[5] B.C.J. Moore. *Perceptual consequences of cochlear damage*. Oxford University Press, 1995.

F1-F2	900	1050	1300	1500	1900	2100	3000
300					y		i
380				ø		e	
500	o	ɔ	œ		ɛ		
650		ɑ	a				

Table 1: Fréquences formantiques des voyelles synthétiques (voix masculine). Pour chaque F1, il existe au moins deux voyelles qui ne diffèrent que par la seule fréquence de F2. À l'inverse, il existe trois paires de voyelles de même F2 mais de F1 différent.

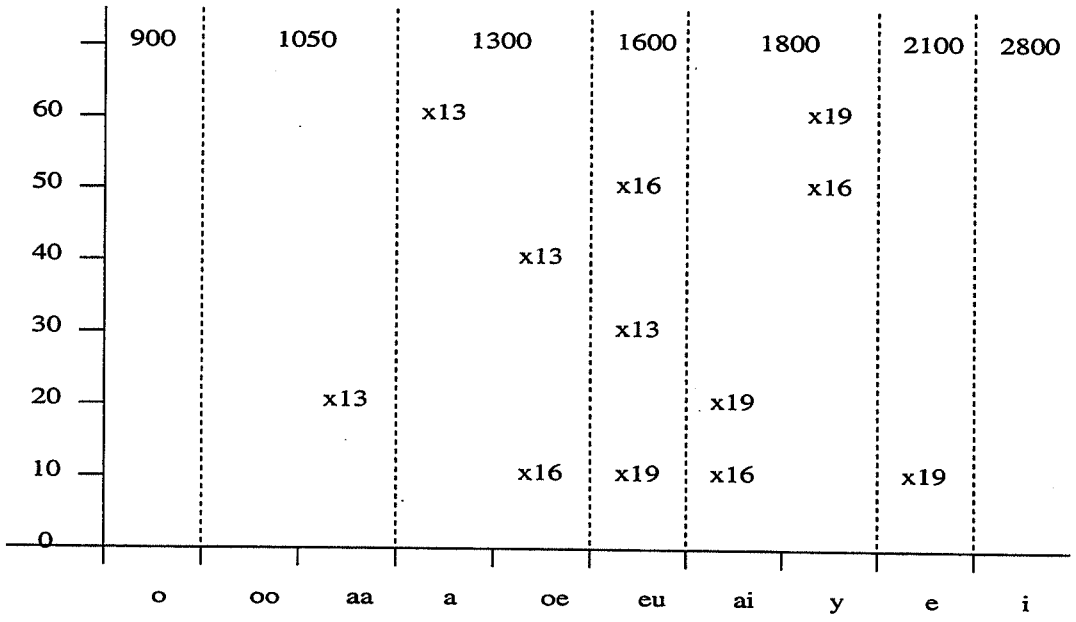


Figure 1: Identification des voyelles après simulation d'une perte d'audibilité locale centrée à 1300 (x13) 1600 (x16) et 1900 Hz (x19). L'axe des ordonnées représente la baisse des taux d'identification des voyelles modifiées par nos simulations par rapport à ceux des voyelles synthétiques non modifiées. Les symboles représentent les voyelles /o,ɔ,ɑ, a,œ,ø,ɛ,y,e,i/. Les lignes verticales en pointillé séparent les voyelles qui possèdent un F2 différent. Les valeurs de celui-ci sont indiquées en haut de la figure.

Dyslexie et déficit du traitement temporel : relation entre Jugement d'ordre et durée des sons de parole

Sonia De Martino, Robert Espesser, Véronique Rey, Michel Habib
Laboratoire Parole et Langage, U.R.A. 261 CNRS & Université de Provence,
29 av. Robert Schuman 13261 Aix en Provence, cedex 1

ABSTRACT

Developmental dyslexia may result from a general, non specific, defect in perceiving rapidly changing auditory signals. 13 phonological dyslexics (age 10-13) and 10 controls matched for chronological age were compared on a Temporal Order Judgment (T.O.J.) task using the succession of two consonants (/p/-/s/) within a cluster. The task included two additional conditions where either the two stimuli were artificially slowed or the interstimulus interval was expanded. As expected, the T.O.J. performance was significantly poorer in dyslexics than in controls. Moreover, in the « slowed speech » condition dyslexics'performance improved, whereas no significant improvement occured when increasing the interstimulus interval.

1. INTRODUCTION

Le déficit phonologique, reconnu comme central dans la dyslexie développementale pourrait résulter d'un déficit général, non spécifique, de la perception des changements acoustiques rapides du signal. Cette hypothèse a soulevé un certain intérêt notamment après les travaux qui ont mis en évidence que la manipulation des caractéristiques temporelles des stimuli acoustiques pouvait améliorer les performances langagières dans certains troubles de la lecture et du langage [Tal96] [Mer96] [Hab99]. Cette théorie repose sur l'observation des faibles performances d'un certain nombre d'enfants ayant des troubles d'apprentissage du langage dans des tâches de Jugement d'Ordre Temporel (J.O.T.). Lorsque l'intervalle entre les stimuli est court (20 à 40 millisecondes) les enfants ont des performances inférieures à celles des contrôles alors que lorsque l'intervalle est long (80 à 120 msec) les performances des enfants ayant des troubles d'apprentissage ne diffèrent plus des sujets contrôles [Tal73] [Tal80]. En dépit d'un certain nombre d'études en faveur de cette théorie (pour une revue : [Far95]), d'autres auteurs ont avancé que le problème n'était pas d'ordre acoustique mais plutôt de nature linguistique [Mod97]. Ces études menées auprès de dyslexiques portent sur des tons et sur des syllabes.

Dans la présente étude, nous nous proposons de tester le J.O.T. des consonnes à l'intérieur d'une séquence syllabique complexe (CCV), il s'agit de deux consonnes /p/ et /s/ au sein d'une structure syllabique CCV courante en français, mais qui représente une difficulté pour les dyslexiques. En effet, les diconsonnes alourdissent la structure syllabique sans pour autant en allonger la durée [Meu94]. Dans le cadre d'un déficit de traitement temporel les diconsonnes illustrent bien la notion d'événements succesifs brefs et les dyslexiques ont tendance à modifier la structure de la syllabe de la diconsonne pour contourner la difficulté en ajoutant un [ə]. Une modification acoustique de la diconsonne devrait améliorer le traitement de ces unités brèves.

Nous émettons deux hypothèses : a) les enfants dyslexiques devraient avoir de meilleurs scores dans l'évaluation de l'ordre des consonnes dans une structure syllabique CCV lorsque la durée des consonnes est allongée (expérimentation 1 et 2), b) dans le cadre de l'hypothèse d'un déficit de traitement temporel, l'allongement de la durée des consonnes devrait être plus pertinente que la simplification de la structure syllabique (passage de la structure CCV à la structure CəCV par l'ajout d'un [ə]) dans la tâche de J.OT).

2. METHODOLOGIE

2.1 Sujets

13 enfants ayant une dyslexie phonologique (11 garçons et 2 filles de 9,8 à 13,7 ans) et 10 enfants contrôles (garçons de 11,5 à 13 ans) appariés sur l'âge chronologique ont participé à cette étude. Les dyslexiques ont été recrutés dans un centre spécialisé dans le traitement et la rééducation de la dyslexie, d'après un QI normal, aucune affection neurologique, aucun trouble auditif ou visuel, aucun déficit attentionnel et une habileté en lecture de deux ans inférieure au niveau attendu (Test de l'Alouette, P. Lefavrais). Les sujets contrôles étaient de jeunes collégiens recrutés en fonction de toute absence de trouble du langage personnel ou familial.

2.2 Expérimentation et procédures

Une étape préliminaire a permis d'évaluer le niveau des performances en conscience phonologique à l'aide de quatre tâches : jugement de rimes, suppression du premier phonème, dictée de non-mots à structure syllabique simple, dictée de non-mots à structure syllabique complexe. Les performances significativement plus faibles des dyslexiques confirment le diagnostic de dyslexie phonologique.

Table 1. Résultats des tests en conscience phonologique

	Jugement de rimes		Dictée non-mots structure complexe	
	Dyslexiques	Contrôles	Dyslexiques	Contrôles
M	81.08	98.18	52.15	96.46
sd	10.57	2.09	22.85	2.81

Table 2. Résultats des tests en conscience phonologique

	Dictée non-mots structure simple		Dictée non-mots structure complexe	
	Dyslexiques	Contrôles	Dyslexiques	Contrôles
M	42.00	99.55	26.62	98.18
sd	26.33	1.61	22.74	1.62

Les groupes de consonnes /ps/ et /sp/ sont deux groupes consonantiques que l'on trouve en français dans les deux ordres. Ces deux groupes ont été insérés dans un contexte vocalique a-a, afin d'éviter des artéfacts acoustiques (troncation de l'explosion du [p], faible audibilité du [s]).

Dans un premier temps, le groupe consonantique d'une durée initiale de 140 millisecondes (msec) (chaque consonne durait 70 msec) a été allongé de 140 msec, donnant ainsi un groupe consonantique de 280 msec. Dans un deuxième temps, on a inséré un [ə] entre les deux consonnes, sans modifier la durée des deux consonnes. On a choisi un [ə], en raison du caractère neutre de cette voyelle et de la faible coarticulation qu'elle présente en contexte consonantique. Quatre non-mots ont été ainsi enregistrés : [apsa] [aspa] [apəsa] [asəpa] dans une chambre anéchoique par une voix masculine naturelle. Les sons ont ensuite été numérisés sur SUN.
Dans les trois expérimentations, les stimuli étaient présentés dans un ordre aléatoire. Les sujets entendaient les non-mots à travers des écouteurs (sous- casque) et devaient désigner l'ordre des deux consonnes en appuyant successivement sur deux touches de l'ordinateur. Afin d'alléger la charge mnésique, les lettres « p » et « s » avaient été dactylographiées sur deux touches du clavier.

Dans la première expérimentation, un bloc de 60 non-mots non modifiés (30 [aspa] et 30 [apsa]) étaient entendus par les sujets, afin de déterminer un éventuel déficit en J.O.T.. Dans une seconde expérimentation, la même procédure était proposée avec 60 non-mots modifiés dans la durée du groupe consonantique (30 [aspa] et 30 [apsa]). Dans la troisième expérimentation, 120 non-mots étaient entendus dont 60 avaient subi une modification temporelle sur les deux consonnes (30 [aspa], 30 [apsa]) et 60 avaient subi une modification par ajout d'un [ə] entre les consonnes (30 [apəsa] et 30 [asəpa]). Les trois expérimentations ont été réalisées sur Macintosh 7200 à l'aide du logiciel Psyscope software.
La passation des trois expérimentations s'effectuait dans un ordre aléatoire.

2.3 Résultats

Expérimentation 1 : Jugement de l'ordre temporel

Cette première condition permet de confirmer que les performances des dyslexiques sont significativement plus faibles que celles des contrôles dans l'habileté à juger l'ordre de deux phonèmes dans un groupe consonantique.

Condition parole normale

	Dyslexiques	**Normolecteurs**
M	61,41	99,24
sd	26,57	1,56

La différence entre les deux groupes est hautement significative (Anova : F=22.082, p=0.0001). Les dyslexiques présentent un déficit dans le jugement de l'ordre de deux consonnes, même lorsque ces consonnes sont phonétiquement très différentes, [p] et [s] diffèrent de plus d'un trait acoustique. Cette observation permet de confirmer la difficulté que représente le groupe consonantique dans une structure CCV en français chez les dyslexiques , relevée comme une erreur typique dans leur lecture et leur transcription écrite.

Expérimentation 2 : Effet de l'allongement de la durée des événements sur le jugement de l'ordre

Condition parole modifiée

	Dyslexiques	**Normolecteurs**
M	81,15	89,39
sd	21,03	25,38

Cette deuxième condition permet de constater l'effet de l'allongement de la durée de la diconsonne sur les performances des dyslexiques dans la tâche de J.OT.

Il y a une interaction groupe x condition significative (Anova à deux facteurs : F=5.46, p=0.0243) qui démontre que l'allongement de la durée des deux consonnes améliore les performances dans le jugement de l'ordre chez les dyslexiques phonologiques. Ces performances sont semblables à celles des normolecteurs.

Expérimentation 3 : Lien entre la durée des événements acoustiques et le jugement d'ordre temporel

Dyslexiques

	Diconsonne modifiée	Ajout d'un [ə]
M	71,28	80,64
sd	22,10	18,44

Normolecteurs

	Diconsonne modifiée	Ajout d'un [ə]
M	81,67	78,33
sd	8,03	7,71

Les résultats de cette expérimentation ne montrent aucune différence significative entre les deux conditions aussi bien chez les dyslexiques que chez les normolecteurs. Ce résultat démontre que l'allongement de la durée des deux consonnes n'est pas un indice plus pertinent pour les dyslexiques, que la simplification de la structure syllabique par l'ajout d'un [ə].

Corrélation entre J.O.T. et déficit phonologique

Un certain nombre de corrélations ont été trouvées entre le taux de réponses correctes des deux premières expérimentations en J.O.T. et certaines tâches de conscience phonologique.
La tâche de suppression du premier phonème et les dictées de non-mots simples et complexes sont positivement corrélées avec les performances en J.O.T. (Suppression premier phonème/J.O.T. CCV non modifié : r=0.644, p=0.0175 ; dictée de non-mots simples et complexes/J.O.T. CCV modifié : r=0.564, p=0.0448).

3. DISCUSSION

Les résultats des expérimentations 1 et 2 confirment la première hypothèse selon laquelle l'allongement de la durée des consonnes améliore les performances des dyslexiques dans l'évaluation de l'ordre de deux consonnes dans une structure syllabique complexe (CCV). Ces résultats sont des arguments en faveur de l'hypothèse d'un déficit de traitement temporel car la difficulté des dyslexiques à traiter des événements brefs comme les syllabes et les tons se retrouvent dans le traitement d'une séquence de deux consonnes.

Les résultats de la troisième expérimentation ne permettent pas de confirmer la deuxième hypothèse selon laquelle l'allongement de la durée des consonnes serait un indice plus pertinent que la simplification syllabique dans la tâche de J.O.T.

Quant à la constatation d'une corrélation entre habiletés métaphonologiques et jugement de l'ordre, elle apporte un argument supplémentaire en faveur du lien présumé entre processus phonologique et traitement temporel, sans permettre cependant d'affirmer la causalité de ce lien.

Cette étude démontre la pertinence de l'allongement de la durée des événement brefs qui peut être un argument en faveur d'une rééducation spécifique, où seul le paramètre acoustique de durée peut être modifié sans porter atteinte à la structure linguistique. Par ailleurs, cette étude ne permet pas de trancher dans le débat théorique entre problème d'ordre acoustique ou problème d'ordre linguistique.

De plus, cette étude montre que l'élaboration de tâches expérimentales est utile et nécessaire dans la recherche d'outils diagnostiques d'un déficit du traitement temporel ou d'outils de prédictibilité en vue d'une méthode de rééducation actuellement proposée dans le traitement des enfants dyslexiques [Tal96] [Mer96] [Hab99].

BIBLIOGRAPHIE

[Far95] Farmer M.E. & Klein R.M. (1995), The evidence for a temporal processing deficit linked to dyslexia : a review, Psychonomic Bulletin & Review, 2 (4), pp 460-493.

[Hab99] Habib M., Espesser R., Rey V., Giraud K., Bruas P., Gres C., (1999), Training dyslexics with acoustically modified speech : evidence of improved phonological performance (abstract), Brain & Cognition, 40 (1), pp143-146.

[Mer96] Merzenich, M.M., Jenkins, W.M., Johnston, P., Schreiner C., Miller S.L. & Tallal P., (1996), Temporal processing deficits of language-learning impaired children ameliorated by training, Science, 271, pp 77-80.

[Meu94] Meunier, C., (1994) Les groupes consonantiques Problématique de la segmentation et variabilité acoustique. Thèse soutenue à l'Université d'Aix en Provence

[Mod97] Mody M., Studdert-Kennedy M., Brady S., (1997), Speech perception deficits in poor readers : auditory processing or phonological coding ?, Journal of Experimental Child Psychology, 64, pp. 199-231.

[Tal80] Tallal P., (1980), Auditory temporal perception, phonics, and reading disabilities in children, Brain & Language, 9, pp. 182- 198.

[Tal96] Tallal P., Miller S.L., Bedi G., Byma G., Wang X., Nagarajan S.S., Schreiner C., Jenkins W.M., Merzenich M.M. (1996), Language comprehension in language-learning impaired children improved with acoustically modified speech, Science, 271, pp. 81-83.

[Tal73] Tallal P., Piercy M., (1973), Defects of non-verbal auditory perception in children with developmental aphasia, Nature, 241, pp. 468-469.

Perception des consonnes occlusives initiales après laryngectomie presque-totale.

E. de Monès (1), S. Hans (1,2), J. Vaissière (2), D. Brasnu (1)
(1) Hôpital Laënnec, Département d'Otorhinolaryngologie et de Chirurgie de la Face et de Cou, 42 rue de Sèvres, 75007 Paris - Université Paris V, UPRESA-CNRS 7018 – Paris, France
Tél. : ++33 (0)144 39 66 58 – Fax : ++33 (0)144 39 66 19
(2) Institut de Phonétique – Université Paris III, UPRESA-CNRS 7018 – Paris, France
Mél : erwan.de.mones@libertysurf.fr

ABSTRACT

Objective : The purpose of this study was to investigate the french stop consonant intelligibility after near-total laryngectomy. Subjects and Methods : Two male speakers were recorded. Corpus were consonant-vowel syllables, with C=/p,t,k,b,d,g/ and V=/a,i/. Perception tests were performed and presented to 12 naïve adult listeners. Listeners' pooled responses were converted to confusion matrices and analyzed for overall intelligibility, voicing and place of articulation features. Results : Overall intelligibility were 81,8 %. In /i/ vowel context, velars consonants /k,g/ were less intelligible (34,2 % and 27,5 % respectively), and perceived as alveolars /t,d/ in 66,3 %. Global voicing confusion score were 6,25 %. Voiced-for-voiceless errors occured more often (10 %).

1. INTRODUCTION

La laryngectomie presque-totale (LPT) a été décrite par Pearson en 1980 [Pea80]. Cette intervention repose sur deux concepts, carcinologique et fonctionnel. Le principe carcinologique est issu d'études histopathologiques des pièces de laryngectomies totales ayant démontré l'absence fréquente d'envahissement tumoral de l'hémilarynx controlatéral à la tumeu. Les résultats carcinologiques ont été validé par de nombreuses publications. Le concept fonctionnel repose sur la réalisation d'un shunt trachéo-laryngo-pharyngé auto-continent permettant la déglutition et la production vocale. Plusieurs études ont confirmé les qualités vocales subjectives des patients opérés de LPT avec un taux de réhabilitation vocale rapporté dans la littérature de 74 % à 85 %. Très peu d'études ont rapporté des paramètres phonatoires objectifs limités à l'analyse de la fréquence fondamentale, au jitter et au shimmer.

La LPT nécessite l'exérèse de la quasi totalité du larynx : seul un aryténoïde est préservé. La reconstruction du pharynx entraîne des modifications anatomiques du tractus vocal. Le but de notre travail était d'étudier les confusions de voisement et les confusions de lieu d'articulation induites par la LPT. Cet article rapporte l'analyse des tests de perrception des consonnes occlusives chez deux sujets masculins opérés par LPT.

2. MATERIEL ET METHODES

Deux sujets masculins atteints d'un carcinome épidermoïde du larynx et traités par laryngectomie presque-totale, âgés de 49 et 63 ans, ont été enregistré 18 mois et 30 mois après l'intervention. Ils étaient indemnes de pathologie tumorale au moment de l'enregistrement.

L'enregistrement acoustique a été réalisé en chambre sourde sur un magnétophone DAT Sony DTC 60FS (Sony, France) avec un microphone Lem EMU 4535 (Lem Communication, Igny, France). Les patients réalisaient une occlusion digitale du trachéostome lors de la phonation.. Le corpus était composé de 12 logatomes du type consonne-voyelle (CV). Les consonnes étaient les 6 consonnes occlusives du français /p, t, k, b, d, g/. Les voyelles étaient la voyelle /a/ et la voyelle /i/. Chaque logatome était répété une dizaine de fois. Les cinq productions les plus proches en terme de ligne mélodique et d'intensité ont été conservées pour les tests de perception, soit 60 stimuli par locuteur.

Les 120 stimuli étaient présentés à l'auditeur par l'intermédiaire d'un ordinateur personnel compatible PC avec le logiciel Sound Forge 3.0 (Sonic Foundry Inc., Madison, Wisconsin). Les stimuli étaient présentés en ordre aléatoire, une seule fois chacun, en raison d'un stimulus toutes les 4 secondes. L'auditeur devait identifier et retranscrire sur une feuille devant lui la consonne perçue parmi les 6 consonnes occlusives du français.

Douze auditeurs (7 hommes et 5 femmes), âgés de 26 à 42 ans (moyenne 32 ans), de langue maternelle française, se sont succédés individuellement. Tous ces auditeurs étaient naïfs et avaient une audition normale.

3. RESULTATS

3.1. Perception globale

Les taux d'identification correcte des 6 consonnes par les 12 auditeurs sont présentés dans la table 1, avec le taux global, le taux pour chacun des locuteurs, et le taux pour chacun des contextes vocaliques.

Le taux général d'identification correcte des 6 consonnes occlusives était de 81,8 % (70,8 % à 89,2 % ; σ = 5,2). Les résultats étaient peu différents pour les deux locuteurs, avec 82,8 % et 78,9 % d'identification correcte respectivement. Pour l'analyse des confusions de voisement et de lieu d'articulation, les résultats des deux locuteurs ont été regroupés.

L'identification des consonnes labiales /p,b/ et alvéolaires /t,d/ était peu modifié par le contexte vocalique /a/ ou /i/.

Ces consonnes étaient correctement identifiées à plus de 80 % (sauf /ta/ à 78,3 %). Les labiales étaient mieux identifiées que les alvéolaires. Les consonnes vélaires /k, g/ étaient nettement moins bien identifiées avec le /i/ qu'avec le /a/ : 34,2 % versus 98,8 % et 27,5 % versus 92,5 % pour /k/ et /g/ respectivement.

Table 1 : Taux d'identification correcte des 6 consonnes, globalement, pour chacun des locuteurs, et dans chaque contexte vocalique

	global	loc. 1	loc. 2	/a/	/i/
p	89,6	91,7	82,5	82,5	96,7
t	83,8	88,3	79,2	78,3	89,2
k	65	71,7	55,8	95,8	34,2
b	98,8	99,2	98,3	99,2	98,3
d	93,8	91,7	96,7	94,2	93,3
g	60	54,2	60,8	92,5	27,5
total	81,8	82,8	78,9	90,4	73,2

loc. 1 et loc.2 : locuteurs 1 et 2.

Les matrices de confusions pour les 6 consonnes sont présentées dans la table 2 pour le /a/ et dans la table 3 pour le /i/. Les résultats sont exprimés en nombre de réponses totales sur un total de 120 stimuli pour chaque syllabe (10 stimuli x 12 auditeurs).

3.2. Perception du trait de voisement

Les taux de confusions de voisement pour chacune des 6 consonnes sont présentés dans la table 4. Le taux global était de 6,25 %. Les confusions de voisement étaient plus fréquentes pour les consonnes non voisées /p, t, k/ que pour les consonnes voisées /b, d, g/. Les confusions de voisement des consonnes non voisées labiale /p/ et alvéolaire /t/ étaient en faveur de la consonne voisée homologue de même lieu d'articulation à 100 % et 95 % respectivement (/p/ percue comme /b/ et /t/ percue comme /d/). Les confusions de voisement de la consonne non voisée vélaire /k/ étaient en faveur d'une consonne de lieu d'articulation différent.

3.3. Perception du lieu d'articulation

Les confusions de lieu d'articulation sont présentées dans les tables 5, 6 et 7. Le taux global de confusions de lieu d'articulation était de 14,4 %. Ces confusions ont surtout affecté les consonnes vélaires /k,g/ dans le contexte vocalique /i/ (64,2 % et 70,8 % respectivement), perçues comme une consonne alvéolaire. Les syllabes /ki/ et /gi/ étaient principalement perçues comme un /ti/ et un /di/ respectivement (table 3 et 4).

4. Discussion

La LPT est une alternative à la laryngectomie totale avec réhabilitation vocale par implant phonatoire (voix trachéo-oesophagienne VTO) [Sin80]. Son intérêt principal repose sur l'absence d'implant qui est un corps étranger à l'origine de complications (détérioration de l'implant, incontinence de la fistule oeso-trachéale. La LPT offre ainsi au patient une qualité de vie supérieure à

Table 2 : matrice de confusion des 6 consonnes dans le contexte vocalique /a/. Nombre de réponses et pourcentage de réponses correctes.

syllabe	b	d	g	k	p	t	taux RC
ba	119			1			99,2
da		113	7				94,2
ga		9	111				92,5
ka	1	1	3	115			95,8
pa	21				99		82,5
ta		8	1	17		94	78,3
résultat voyelle /a/							90,4

RC : réponses correctes.

Table 3 : matrice de confusion des 6 consonnes dans le contexte vocalique /i/. Nombre de réponses et pourcentage de réponses correctes.

syllabe	b	d	g	k	p	t	taux RC
bi	118	1			1		98,3
di	2	112	1		3	2	93,3
gi		76	33	2	1	8	27,5
ki		19	2	41	2	56	34,2
pi	4				116		96,7
ti		12			1	107	89,2
résultat voyelle /i/							73,2

RC : réponses correctes.

Table 4 : Pourcentages de confusions de voisement pour chaque consonne, globalement et dans chaque contexte vocalique.(proportions de confusions homologues : p→b, t→d, k→g, et réciproquement).

	global	/a/	/i/
p	10 (100)	17,5 (100)	3,3 (100)
t	9 (95)	7,5 (89)	10 (100)
k	11 (19)	4,2 (60)	16,7 (10)
b	1	0,8	0,8
d	2	0	4,2 (40)
g	5	0	9,2 (18)
total	6,25	5	7,4

Table 5 : Taux des confusions de lieu d'articulation pour chacune des 6 consonnes, globalement et dans les deux contextes vocaliques /a/ et /i/.

	confusions de lieu d'articulation		
	global	voyelle /a/	voyelle /i/
p	0	0	0
t	8	15	0,8
k	33	1,7	64,2
b	0,5	0,8	0,8
d	5,4	5,8	5,0
g	39	7,5	70,8
total	14,4	5,1	23,6

la laryngectomie totale avec réhabilitation par implant phonatoire. L'intégration récente de cette intervention comme une option thérapeutique par des équipes chirurgicales de tous les continents atteste de son intérêt. La création du shunt trachéo-laryngo-pharyngé représente un modèle de production de voix pathologique. Les études publiées dans la littérature ont rapporté soit une analyse subjective de la production vocale soit des

Table 6 : Matrice des confusions de lieu d'articulation, en pourcentage, dans le contexte vocalique /a/.

voyelle /a/	réponse		
	pb	td	kg
labiales /p,b/	99,6		0,4
alvéolaires /t,d/		89,6	10,4
vélaires /k,g/	0,4	4,2	95,4

Table 7 : Matrice des confusions de lieu d'articulation, en pourcentage, dans le contexte vocalique /i/.

voyelle /i/	réponse		
	pb	td	kg
labiales /p,b/	99,6	0,4	
alvéolaires /t,d/	2,5	97,1	0,4
vélaires /k,g/	1,3	66,3	32,5

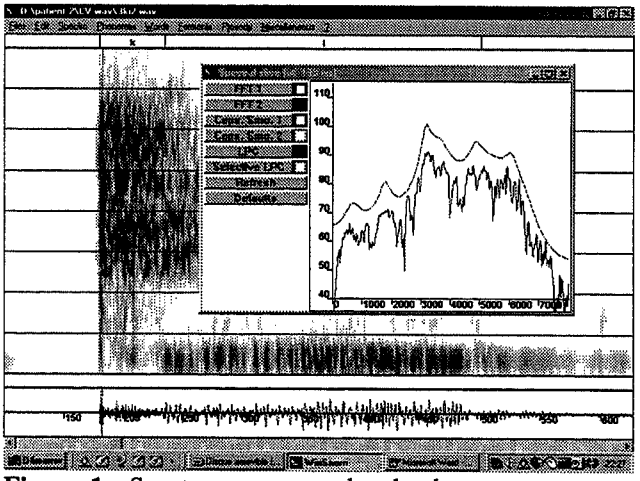


Figure 1 : Spectrogramme en bandes larges et coupe spertrale du burst du /ki/ prononcé par le locuteur 2, et identifié 9 fois comme /t/ et 3 fois comme /k/ par les 12 auditeurs.

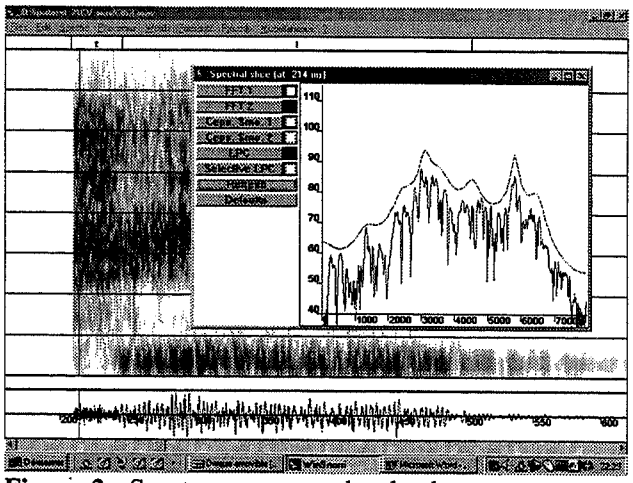


Figure 2 : Spectrogramme en bandes larges et coupe spectrale du burst d'un /ti/ prononcé par le locuteur 2, et correctement identifié par les 12 auditeurs.

paramètres acoustiques fréquentiels. Notre travail repose sur deux hypothèses induites par la résection de la quasi totalité du larynx et les modifications du tractus vocal par la reconstruction : l'apparition de confusions de voisement et d'articulation.

Le choix des logatomes CV par rapport à des mots permet d'éviter le biais de perception lié à la signification du mot. Dans le logatome CVCVCVC, seul la consonne initiale a été étudiée pour les raisons suivantes : (i) En français, les consonnes initiales sont les plus résistantes. Ainsi, Bourciez et Bourciez [Bou95] ont montré que l'évolution des mots du latin au français s'accompagnait d'une disparition des consonnes finales et de la conservation des initiales (exemple : campus > campu > camp). (ii) En français, pour un sujet normal, il est plus facile de continuer le voisement, en intervocalique, que de l'arrêter (exemple : abeille/apiculteur), et il existe une tendance à anticiper la pause, c'est à dire à devoiser en fin de mot (exemple : vive/vif) [Vai97]. Le choix des voyelles /a/ et /i/ a été dicté par leur fréquence en français.

Les résultats des tests de perception étaient très proches en terme d'intelligibilité globale et de confusions pour chacune des consonnes (Table 1). Les 6 consonnes occlusives initiales ont été globalement bien identifiées (81,8 % au total, 70,8 % à 89,2 %, σ = 5,2). Ce taux est concordant avec les résultats rapportés dans la littérature : la voix était considérée comme « bonne » chez 74 % à 85 % des patients. Dans la voix normale, Bonneau et al ont rapporté un taux moyen de 89 % pour les trois contextes vocaliques /a,i,u/ chez 5 locuteurs [Bon96]. Nos résultats ont montré que le trait de voisement était bien réalisé dans la LPT. Les confusions de voisement étaient de 6,25 % au total. Elles affectaient surtout les consonnes non voisées, perçues dans 10 % comme voisées (Table 5). Les consonnes /p/ et /t/ étaient alors principalement confondues avec leurs homologues voisées de même lieu d'articulation, /b/ et /d/ respectivement, dans les deux contextes vocaliques (89 % à 100 % de confusions homologues) (Table 4). La consonne vélaire /k/ associait plus fréquemment une confusion de voisement et une confusion de lieu d'articulation, surtout dans le contexte vocalique /i/ (10 % de confusions homologues).

Le contexte vocalique modifiait l'intelligibilité de certaines consonnes. Les 6 consonnes étaient aisément identifiées dans le contexte vocalique /a/. Dans le contexte vocalique /i/, les consonnes vélaires /k,g/ étaient mal identifiées (Table 1). Stevens a montré que la perception du lieu d'articulation des consonnes initiales dépendait du spectre du burst [Ste97]. On sait aussi que la fréquence centrale du bruit de relâchement de la consonne vélaire dépend fortement de la voyelle : il est grave devant les voyelles postérieures, moyen devant la voyelle /a/ et plus aigu devant les voyelles antérieures. Stevens a montré que cette fréquence correspondait à la résonance de la cavité antérieure. Or la voyelle /i/ est plus antérieure que la voyelle /a/ : le lieu d'articulation de /k/ et /g/ devient palatal, d'où un rapprochement de l'articulation de /t/ et /d/. La cavité antérieure se raccourcit, augmentant

la fréquence centrale du burst de /ki/ et /gi/, et sa largeur de bande. Le burst de /ki/ et /gi/ est alors proche du burst de /ti/ et /di/, comme les figures 1 et 2 l'illustre. Il faut une plus grande précision articulatoire pour rendre plus pertinente la différence acoustique. De même, Vaissière [Vai97] a montré que la consonne /g/ était l'occlusive voisée la moins fréquente en langue française. Cette interaction entre la consonne /k/ et la voyelle /i/ a été bien analysée dans la voix normale [Kew83, Blu79, Bon96]. En français, Bonneau et al ont rapporté un taux d'intelligibilité de 78 % pour /ki/ contre 90 % et 98 % pour /ka/ et /ku/ respectivement [Bon96]. Dans notre étude, le taux des vélaires dans le contexte vocalique /i/ était néanmoins beaucoup plus mauvais que dans la voix normale (34,2 % pour /ki/ et 27,5 % pour /gi/).

Nous ne pouvons pas expliquer la fréquence plus importante des confusions de voisement dans le sens sourde/sonore que dans le sens sonore/sourde. Le mécanisme physiologique "actif" de contrôle on/off de la mise en vibration ou de la non-mise en vibration pourrait avoir plusieurs origines. (i) selon Ladefoged [Lad83] et Stevens [Ste91], chez les sujets normaux, le dévoisement correspondrait à un état de tension musculaire, et le voisement aurait tendance à réaliser une "économie". Cet état de tension serait du à une élévation du larynx qui met en tension les cordes vocales, associé à un rétrécissement des cavités pharyngées. (ii) Pour les consonnes occlusives non voisées, Hirose [Hir77] a montré le rôle actif du muscle crico-aryténoïdien postérieur dans l'ouverture postérieure de la glotte. Cependant toutes ces structures sont modifiées dans la LPT.

Les confusions d'articulation induite par l'intervention sont certainement lié à une augmentation de la dépendance entre la source et les articulateurs, notament la base de la langue. Notre étude a montre une interrelation entre les confusions de voisement et d'articulation pour les velaires, certainement en rapport avec des phénomènes de compensation entre les structures du tractus vocal. Des études complémentaires par vidéofibroscopie avec stroboscopie pourraient permettre une analyse morphodynamique des structures restantes et objectiver la dépendance source-articulateurs.

Notre travail, en analysant la perception des consonnes occlusives initiales, a montré que i) le trait de voisement était réalisé et perçus dans plus de 80 % des cas ii) les confusions de lieu d'articulation plus fréquentes que dans la voix normale pour les consonnes vélaires avec le /i/ traduisent des phénomènes de coarticulation majorés, probablement liés au geste chirurgical.

En pratique clinique ces constatations pourraient orienter la rééducation de ces patients en utilisant des stratégies de compensation permettant une meilleure intelligibilité de la production de la parole.

BIBLIOGRAPHIE

[Blu79] Blumstein S.E. (1979), « Acoustic invariance in speech production : evidence from measurements of spectral characteristics of stop consonants », J. Acoust. Soc. Am., Vol. 66, pp. 1001-1017.

[Bon79] Bonneau A. (1979), « Perception of the place of articulation of french stop bursts », J. Acoust. Soc. Am, Vol. 100, pp. 555-564.

[Bou95] Bourciez E. (1995), Phonétique Française, Klincksieck.

[Doy89] Doyle P.C. (1989), « Perception of pre-vocalic and post-vocalic consonants produced by tracheoesophageal speakers », J. Otolaryngol., Vol. 18, pp. 350-353.

[Gom89] Gomyo Y. (1989), « Perception of stop consonants produced by esophageal and tracheoesphageal speakers », J. Otolaryngol., Vol. 18, pp. 184-188.

[Hir77] Hirose H. (1977), « Laryngeal adjustments in consonant production », Phonetica, Vol. 34, pp. 140-64.

[Kew83] Kewley-Port D. (1983), « Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants », J. Acoust. Soc. Am., Vol. 73, pp. 1779-1793.

[Lad83] Ladefoged P. (1983), The linguistic use of different phonation types, Vocal Fold Physiology : contemporary research and clinical issues. San Diego, College Hill.

[Pea80] Pearson B.W. (1980), « Extended hemilaryngectomy for T3 glottic carcinoma with preservation of speech and swallowing », Laryngoscope, Vol. 90, pp. 950-961.

[Sin80] Singer M.I. (1980), « An endoscopic technique for restoration of voice after laryngectomy », Ann. Otol. Rhinol. Laryngol., Vol. 89, pp. 529-533.

[Ste81] Stevens K.N. (1981), The search for invariant acoustic correlates of phonetic features, Perspectives on the study of speech, Erlbaum, Hillsdale.

[Ste91] Stevens K.N. (1991), Vocal fold vibration for obstruent consonants, Vocal Fold Physiology : acoustic, perceptual and physiological aspects of voice mechanisms, Singular Publishing Group.

[Ste97] Stevens K.N. (1997), « Articulatory-acoustic-auditory relationship », The handkook of phonetic sciences, Blackwell Publishers Ltd, Oxford

[Vai97] Vaissière J.(1997), Phonological use of the larynx, Proceedings larynx, Marseille.

Modèles pour l'intégration de représentations perceptives et motrices dans l'acquisition du langage

Jean-Luc Schwartz, Louis-Jean Boë, Yanick Paviot

Institut de la Communication Parlée - CNRS / INPG / Université Stendhal
INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex 1
Tél.: ++33 (0)476 57 47 12 - Fax: ++33 (0)476 57 47 10
Mél: schwartz@icp.inpg.fr - http://www.icp.inpg.fr

ABSTRACT

We propose a view of speech perception in which a basic goal of perceptual representations is to enable action control, and we show that in this framework, speech perceptual and motor representations are linked by bilateral constraints. Then we introduce a model of speech acquisition in which sensori-motor representations emerge dynamically in the course of babbling and learning, together with a preliminary experiment displaying its behavior in a simple case. At last, we present a new model for predicting palatal contacts from articulatory commands, and we discuss its interest for speech control.

1. INTRODUCTION : UN POINT DE VUE CRITIQUE SUR LES THEORIES DE LA PERCEPTION DE LA PAROLE

Les travaux sur la perception de la parole sont en grande partie structurés par un débat désormais classique. *Pour les tenants des théories "auditives"*, la perception de la parole est l'ensemble des mécanismes permettant au système auditif et au cerveau de décoder le message émis par l'interlocuteur, et inscrit dans le signal acoustique. Il s'agit donc de comprendre comment fonctionne le traitement biologique de l'information dans le système auditif, indépendamment des mécanismes de production. *Pour les tenants des théories "motrices"*, la parole n'est "intelligible" – pour l'auditeur comme pour le chercheur – que dans l'organisation des gestes articulatoires, et la perception de la parole est l'ensemble des mécanismes permettant à l'audition (ou aux autres sens concernés) de retrouver ces gestes qui structurent la matière sonore. Ces mécanismes, qu'ils soient proposés sous une forme qualitative ou calculatoire, ne font en général pas appel à des connaissances particulières sur le système auditif.

Entre théories auditives, qui n'exploitent aucune donnée sur l'action, et théories motrices, qui n'en exploitent pas plus sur l'audition, est apparue progressivement une troisième voie [Lin90]. Elle installe l'*interaction locuteur-auditeur* au cœur de la structuration du langage. Si ce programme de recherche met la notion *d'information* au centre du raisonnement, il faut bien reconnaître qu'il n'a proposé rien de concret pour mesurer, modéliser cette information.. La raison en est d'après nous que cette théorie reste au niveau de *la surface de l'interaction* – le son, quantifié par l'information qu'il porte – d'où la difficulté d'avancer sur la voie d'une mise en œuvre quantitative. La piste que nous cherchons à explorer ici consiste à remettre en jeu les représentations auditives et motrices, au sein d'une *Théorie de la Perception pour le Contrôle de l'Action*. Nous allons d'abord présenter les principes de cette théorie, puis décrire des outils de modélisation développés pour expérimenter, dans ce cadre, sur l'émergence conjointe et contrainte de représentations sensorielles et motrices des objets phonétiques.

2. UNE THEORIE DE LA PERCEPTION POUR LE CONTROLE DE L'ACTION (TPCA)

Dans ce cadre théorique, la perception de la parole est conçue comme l'ensemble des processus perceptifs (auditifs, visuels, somesthésiques) et des représentations associées permettant le contrôle de ses propres actions (leur spécification), le suivi des actions de l'autre (leur récupération via leur mise en forme) et contribuant à la morphogenèse des unités du langage (en fournissant des contraintes et pressions périphériques sur l'émergence d'une phonologie).

Le *suivi auditif des gestes articulatoires* passe par la capture et la caractérisation de leurs deux composantes de base, les cibles et le phasage. Le système auditif sait mesurer événements temporels et caractéristiques spectrales, à partir desquels il est possible de remonter à certaines caractéristiques majeures des gestes, telles que voisement des plosives, lieu d'articulation des voyelles ou des plosives en contexte.

Mais *l'audition "met en forme" les gestes*. D'abord, parce que cette représentation auditive du spectre ne permet qu'une caractérisation partielle du contrôle. Il existe des gestes "non audibles", et donc non caractérisables ou récupérables par les traitements auditifs. C'est précisément ce mécanisme qui est à la base de la coarticulation. Ensuite, parce que l'audition structure le geste [Ste89]. C'est ce qui fait la difficulté de caractériser un geste phonologiquement pertinent, indépendamment de toute capacité perceptive. Cette capacité qu'a le système perceptif de structurer l'espace des gestes peut permettre d'y loger de la phonologie (voir l'exemple du contraste [i]-[y], [Sch93]). Et puis, en sens inverse, *l'action contraint les percepts*. L'information auditive est parfois incomplète, et alors doivent rentrer en compte des procédures de régularisation, comme c'est le cas pour la perception visuelle (voir par exemple [Bai00] ou

[Loe97]). On en arrive donc à un format de représentation qui est bien *intrinsèquement sensori-moteur*, représentations pour *spécifier et récupérer des contrôles*, et qui ne sont ni purs produits sensoriels, ni purs objets moteurs inférés, mais des percepts multimodaux régularisés par l'action, ou des gestes remis en forme par la perception multimodale. Pour l'essentiel, les données disponibles sur la perception de la parole montrent d'ailleurs que l'auditeur perçoit comme il agit ou peut agir, et agit selon ce qu'il perçoit [Sch98].

3. UN MODELE D'ACQUISITION CONJOINTE DE REPRESENTATIONS SENSORI-MOTRICES

La TPCA doit s'instancier par un *agent de communication multisensoriel* : un androïde disposant de la capacité d'articuler, d'entendre, de voir, de ressentir par le toucher et la proprioception l'état de son conduit vocal et celui de son interlocuteur. C'est cet agent qui permettra de déposer en un même réceptacle nos connaissances et nos hypothèses sur la perception et la production de parole, de les mettre dans un champ évolutif correspondant à la croissance et à l'acquisition de la parole, puis éventuellement à ses conditions d'émergence et d'évolution. Cet androïde est constitué d'un modèle articulatoire ([Mae90], avec son environnement informatique [Boë95]), pourvu de capteurs sensoriels : l'extraction de formants pour l'ouïe, l'extraction de paramètres labiaux pour la vue, la somesthésie, incluant contacts labiaux et palataux, sur laquelle nous reviendrons, et la proprioception, que l'on peut simuler par la connaissance des paramètres articulatoires de commande. Il doit alors être lancé dans l'apprentissage de ses représentations sensori-motrices. Quelques modèles d'apprentissage de *relations* sensori-motrices ont été proposés dans la littérature ces dernières années [Gue95], [Bai97]. Or, ces modèles ont en commun de supposer acquises les données d'un espace auditif et d'un ensemble de commandes articulatoires *préexistants*, puis de se doter d'une phase, dite de babillage, lors de laquelle le système explore *de manière systématique* ses capacités motrices et apprend donc l'ensemble de la fonction articulatori-acoustique sur tout l'espace d'entrée. Au contraire, un point central pour nous est l'hypothèse selon laquelle les représentations perceptives et motrices ne sont pas préexistantes, mais au contraire émergent conjointement, en se contraignant mutuellement, au cours d'un mécanisme d'exploration *progressive* de l'espace des phases articulatori-auditives.

Prenons l'exemple de [u]. Les travaux sur le modèle articulatoire [Boë92] ont montré qu'il existait pour cette voyelle deux contrôles possibles, l'un vélo-palatal [u] et l'autre vélo-pharyngal [u]*. Or, l'un seulement de ces deux contrôles est mis en œuvre par les locuteurs français, l'articulation vélo-palatale. La raison en est-elle une impossibilité articulatoire, ou un coût exhorbitant, pour le geste vélo-pharyngal [u]* ? Probablement pas, puisque c'est un contrôle du même type qui est à la base du [o]. On peut alors être amené à supposer que c'est la configuration atteinte naturellement à partir de la configuration neutre qui est choisie, postérieurement à la phase de babillage qui déclenche les processus d'acquisition du langage. Cette mémoire du premier *mapping articulatori-acoustique* serait ensuite préservée dans la représentation du [u] chez l'adulte [Abr96], comme semblent le montrer les blocages articulatoires sur la configuration vélo-palatale observés dans des tâches de perturbation [Sav95], dans lesquelles le recul vers une articulation vélo-pharyngale pourrait pourtant être perceptivement rentable.

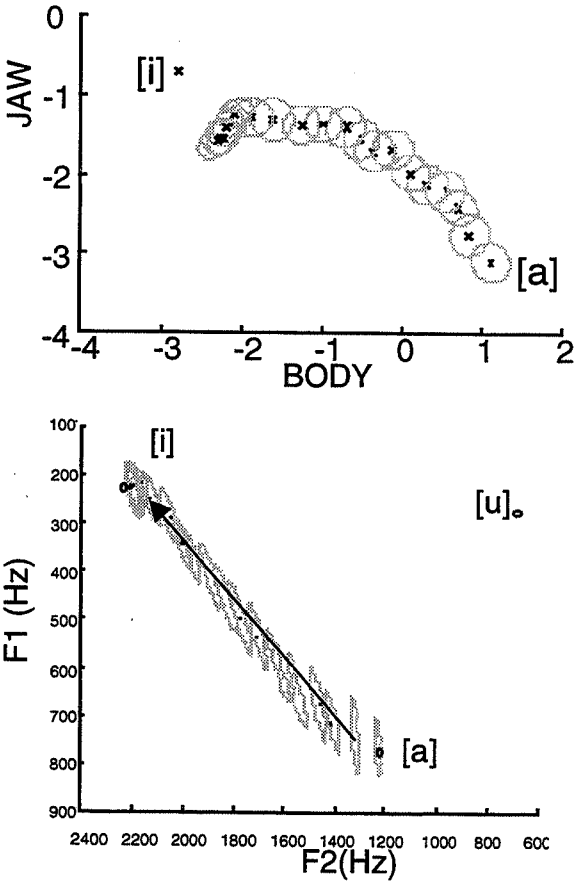


Figure 1 – Apprentissage par exploration sensori-motrice

On part d'une exploration locale autour de [a], puis on se déplace vers [i], en apprenant de proche en proche la correspondance articulatori-acoustique. On observe la cohérence de la voie articulatoire découverte: avancée du corps de la langue, montée de la mâchoire.

L'apprentissage se fait, dans notre modèle, en deux temps : d'abord, un processus endogène, celui du babillage canonique, consistant simplement à explorer les relations sensori-motrices par des gestes d'ouverture-fermeture de type [bəbəbə] ou [djədjədjə] ; puis un processus exogène, celui de l'apprentissage de la langue maternelle. Un ingrédient crucial de notre androïde doit donc être la capacité d'apprendre d'abord des *correspondances statistiques locales entre gestes et percepts multisensoriels* au cours du babillage endogène, puis de tenter *d'extrapoler* ces relations pour *explorer et apprendre de nouveaux chemins*, et s'approcher petit à petit de ses cibles. Ce principe théorique a été introduit et

exploité avec succès dans le domaine de la robotique cognitive, pour l'apprentissage de comportements moteurs élémentaires puis plus complexes [Bes98]. Nous présentons sur la Fig. 1 un résultat préliminaire.

Nous lançons d'abord notre modèle dans une activité motrice stéréotypée, consistant à produire une configuration vocalique de type [a], qui fournit pour l'instant une version simplifiée (voyelle centrale ouverte) de la production des gestes d'ouverture-fermeture précités. Cette activité motrice est réalisée avec un bruit faible sur chaque articulateur, et le système peut alors « apprendre » la relation sensori-motrice autour de la configuration canonique. L'apprentissage va consister à observer l'ensemble des vecteurs (Mi, Fj), où les Mi sont les 7 paramètres articulatoires (mâchoire, corps, dos, pointe de la langue, ouverture et protrusion des lèvres, élévation du larynx), variant faiblement autour de la cible [a], et les Fj sont les 3 premiers formants du son résultant. Cet ensemble est modélisé par une gaussienne (donc par une moyenne et une matrice de covariance à 10 dimensions). Le système cherche alors, à partir de ce noyau d'apprentissage local, à explorer son espace articulatori-acoustique pour atteindre une cible supposée fournie par l'environnement, ici la configuration [i]. Mathématiquement, l'apprentissage fournit une relation probabiliste reliant entrée articulatoire (M) et sortie acoustique (F) par la loi de probabilité p(M, F), et l'exploration consiste à maximiser la probabilité p(M/F) pour un vecteur (F) cible donné. L'évolution se fait de proche en proche : à chaque pas de calcul, on explore une région de l'espace articulatori-acoustique qui permet de se rapprocher de la cible, et la nouvelle matrice de covariance apprise permet de se guider efficacement vers l'objectif à atteindre. On peut ainsi découvrir pas à pas les principaux degrés de libertés articulatoire, et apprendre sous une forme statistique les relations entrée-sortie. Ce principe d'apprentissage sera peu à peu étendu à des simulations plus complexes et plus réalistes, dans le but de mieux comprendre comment peuvent ainsi s'élaborer conjointement représentations sensorielles et motrices. Il devra intégrer également les données d'autres capteurs, comme celui que nous allons présenter maintenant.

4. UN MODELE DE CAPTEUR OROSENSORIEL

Les différents modèles d'apprentissage et de production de la parole supposent pour la plupart l'existence d'informations orosensorielles, notamment sur la configuration des lieux de constriction maximale dans le conduit vocal. Un certain nombre de données expérimentales confirment la nécessité de telles informations pour le contrôle de la parole. Or, il n'existe à notre connaissance aucun modèle de capteur tactile susceptible de fournir des données orosensorielles pour le contrôle de l'action en production de la parole. Nous avons donc décidé d'élaborer un premier modèle, très simple, capable de prédire la forme des contacts palataux en fonction des commandes articulatoires. Le point de départ est une série de données palatographiques sur les voyelles et les plosives en contexte vocalique [Rec91], [Rec93]. Ces données sont caractérisées par 5 paramètres {L1..L5} correspondant au nombre de points de contacts à partir de l'arrière du palais, en cinq positions, de la périphérie du palais (au voisinage des dents) vers la ligne médiane. A chaque configuration phonétique disponible dans les données publiées (les voyelles [i e ɛ u o ɔ a] ; et les plosives [t] en contexte symétrique [u], [i], [a], et [k] en contexte symétrique [i], [a]), nous avons associé une configuration articulatoire prototypique sur le modèle [Ber94], [Val95]. Puis nous avons défini un associateur simple permettant de passer des 4 paramètres articulatoires que nous avons considérés pertinents {P1..P4} (la mâchoire, le corps, le dos et la pointe de la langue) aux cinq paramètres décrivant le contact palatal. Cet associateur est linéaire à seuil, défini par l'équation :

$$Li = f(\sum wijPj + wi0)$$

où wij et wi0 sont les poids et les biais à apprendre, et f est la fonction seuil telle que :

$$f(x) = x \text{ si } x > 0 \; ; \; f(x) = 0 \text{ si } x \leq 0$$

Le nombre de données permettant l'apprentissage est bien sûr faible (pour chaque paramètre Li à prédire, nous disposons de 12 configurations pour calculer 5 paramètres). Cependant, les configurations disponibles présentent l'intérêt majeur de relativement bien "couvrir" l'espace des configurations palatales possibles (à l'exception des latérales). Aussi, notre modèle s'avère efficace, tant pour prédire effectivement toutes les données d'apprentissage (Fig. 2), que pour généraliser à des configurations non apprises. Nous avons étudié l'ensemble des configurations de contacts palataux prédites lorsque l'on explore un grand nombre de configurations articulatoires permettant d'atteindre des valeurs formantiques autour des cibles [i], [a], [u] (Fig. 3a). Ces configurations produisent des contacts palataux compatibles avec la variabilité observée pour des productions de ces voyelles dans différents contextes consonantiques [Rec91] (Fig. 3b). On observe au passage que ce modèle de "capteur palatal" fournit une information cohérente avec l'information auditive, ce qui pourrait fournir un ingrédient essentiel au contrôle en situation de perturbation.

CONCLUSION

Ce travail préliminaire introduit les ingrédients d'un programme de modélisation et d'expérimentation sur le développement conjoint de représentations sensorielles et motrices co-structurées. Il s'agit d'un premier pas dans le chemin de la synthèse que nous proposons entre théories auditives et motrices. L'interaction locuteur-auditeur y est placée au cœur de l'approche, non pas diluée dans une théorie informative "faible", mais *resituée dans le parcours dynamique de l'exploration sensori-motrice qui prend place dans l'ontogenèse*. Après une étape de validation, nous mettrons en relation cette dynamique

avec celle de la phylogenèse, pour laquelle nous avons déjà avancé une hypothèse théorique [Sch97].

Remerciements – Ce travail a bénéficié de discussions fécondes et de l'aide de C. Abry, D. Beautemps, P. Bessière, R. Laboissière et O. Lebeltel.

BIBLIOGRAPHIE

[Abr96] Abry, C., Badin, P. (1996), "Speech Mapping ...", *Proc. 4th Speech Prod. Sem.*, Autrans, 175-184.

[Bai97] Bailly, G. (1997), "Learning to speak. Sensory-motor control of speech movements" *Speech Com., 22*, 251-267.

[Bai00] Bailly, G. (2000), *Représentations phonétiques et technologies vocales*, HdR, INP Grenoble.

[Ber94] Berrah A.R. (1994), *"L'émergence des structures sonores : les syllabes consonnes/voyelles"*, DEA Sciences Cognitives, INP Grenoble.

[Bes98] Bessière, P. et al. (1998), "Proposition pour une théorie probabiliste des systèmes cognitifs sensori-moteurs", *Intellectica, 26-27*, 257-311.

[Boë92] Boë, L.J. et al. (1992), "The geometric vocal tract variables controlled for vowel production", *J. Phon., 20*, 27-38.

[Boë95] Boë, L.J. et al. (1995), "Vers une unification des espaces vocaliques", in C. Sorin et al. (eds.) *Levels in Speech Communication: Relations and Interactions* (pp. 63-71). Elsevier B.V.

[Gue95] Guenther, F.H. (1995), "Speech sound acquisition, coarticulation, and rate effects in a neural model of speech production", *Psychological Review, 102*, 594-621.

[Lin90] Lindblom, B. (1990), "On the notion of possible speech sound", *J. Phon., 18*, 135-152.

[Loe96] Lœvenbruck, H., Perrier, P. (1997), "Motor control information recovering from the dynamics with the EP hypothesis", *Proc. Eurospeech'97, 4*, 2035-2038.

[Mae90] Maeda, S. (1990), "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal - tract shapes using an articulatory model", in W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Modellling* (pp. 131-149), Kluwer.

[Rec91] Recasens, D. (1991), "An electropalatographic and acoustic study of consonant-to-vowel coarticulation", *J. Phon., 19*, 177-192.

[Rec93] Recasens, D. et al. (1993), "An electropalatographic study of stop consonant clusters", *Speech Comm., 12*, 335-356.

[Sav95] Savariaux, C. et al. (1995), "Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production", *JASA, 98*, 2428-2442.

[Sch93] Schwartz, J.L. et al. (1993), "Interindividual and cross-linguistic strategies for the production of the [i] vs [y] contrast", *J. Phon., 21*, 411-425.

[Sch97] Schwartz, J.L. et al. (1997), "The dispersion-focalization theory of vowel systems", *J. Phon., 25*, 255-286.

[Sch98] Schwartz, J.L., Abry, C. (1998), "La théorie motrice de la perception de la parole", *Sciences et Vie, 204*, 114-113.

[Ste89] Stevens, K.N. (1989), "On the quantal nature of speech", *J. Phon., 17*, 3-45.

[Val95] Vallée N. et al. (1995), "Vowel prototypes for UPSID's phonemes", *Proc. XIIIth Int. Congr. of Phonetic Sciences, 1*, 424-427.

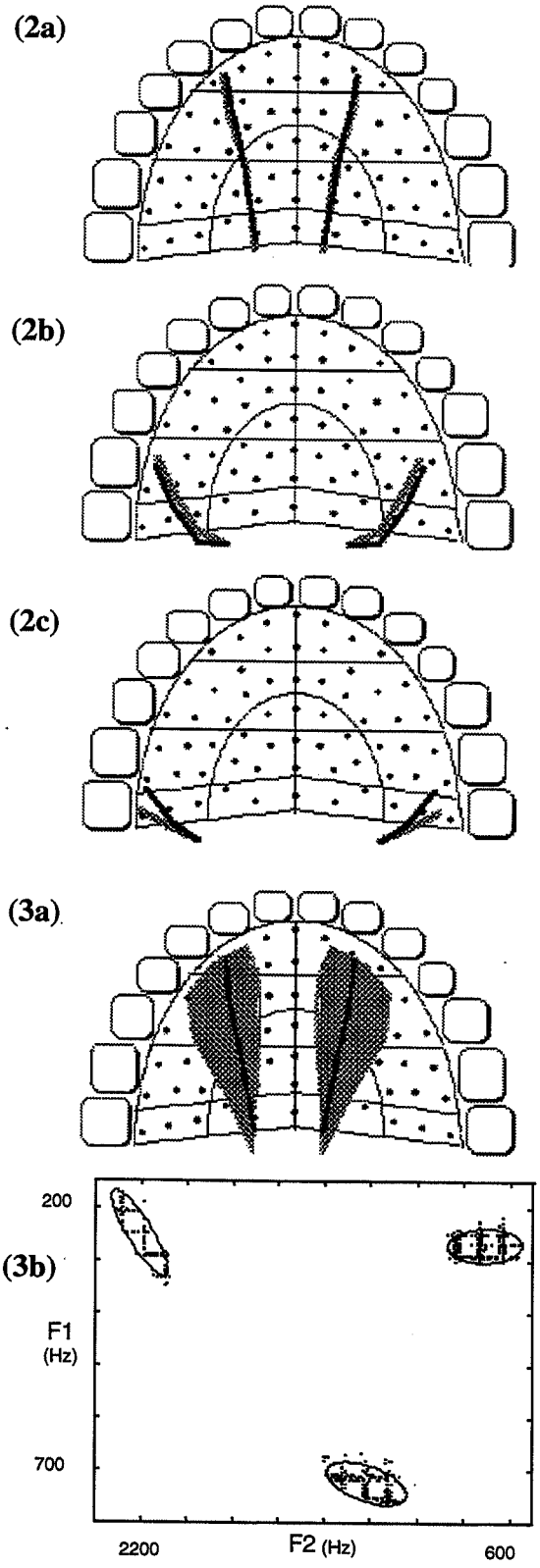


Figure 2 – Configurations prédites (en noir) et observées (en gris) des contacts palataux pour :
(a) [i] ; (b) [a] ; (c) [o].

Figure 3 – Configurations prédites pour un millier de configurations articulatoires autour de [i] :
(a) contacts prédits ; (b) variations acoustiques

Particularités articulatoires de la dyslexie développementale phonologique

Muriel Lalain, Didier Demolin, Michel Habib, Noël Nguyen, Bernard Teston

Laboratoire parole et langage, CNRS ESA 6057
Université de Provence , Aix-en-Provence
muriel.lalain@lpl.univ-aix.fr

ABSTRACT

This study is concerned with the articulatory and acoustic characteristics of speech in dyslexic children. We asked to what extent developmental phonological dyslexia is associated with potential disorders in the fine-grained control of articulatory movements in speech. Aerodynamic and acoustic data were simultaneously gathered in a reading task and a sentence-repetition task for 10 dyslexic children and 2 groups of control children. The results revealed specific segmental error patterns, as well as subtle differences in the articulatory/acoustic structure of stops, for the dyslexic children compared to the control groups. Implications for the potential origin of developmental phonological dyslexia are discussed.

1 INTRODUCTION

La dyslexie, trouble de l'apprentissage du langage écrit touche environ 10% de la population. Aujourd'hui, la possibilité d'attribuer au trouble une origine neurologique est communément admise : les travaux en neurologie et dans le domaine de la perception de la parole font état de particularités aux niveaux neuroanatomique (latéralisation hémisphérique, taille du corps calleux, ectopies) et neurofonctionnels (déficits visuel, temporel, de conscience phonologique) [Hab98]. Les quelques travaux menés en production mettent en évidence les relations entre Dyslexie et langage oral, mais révèlent surtout le rôle primordial des mécanismes articulatoires dans la genèse du trouble[Ale91].

Cette évaluation des capacités articulatoires de dix enfants atteints de dyslexie développementale phonologique, lors de l'exercice de lecture et lors d'une tâche de répétition tente de préciser l'implication de ces mécanismes dans la dyslexie de développement. Menée auprès de trois groupes de sujets, cette analyse comparative a ainsi permis d'obtenir une description, à partir de données acoustiques et aérodynamiques, des profils articulatoires des sujets dyslexiques et contrôles, qui révèlent des particularités caractéristiques du trouble.

2 METHODOLOGIE

2.1 Sujets

Trois groupes de sujets ont participé à cette étude comparative : un groupe de sujets dyslexiques (D) et deux groupes de sujets contrôles (T1 et T2) respectivement appariés au premier sur la base de l'âge de lecture (D-T1) et sur la base de l'âge chronologique (D-T2). Chacun des trois groupes est constitué de dix enfants ; les enfants du groupe Dyslexique sont âgés de 10 ans 7 mois à 13 ans 10 mois mais ont un âge de lecture qui correspond à l'âge chronologique des enfants du groupe T1, c'est-à-dire 7 à 8 ans . Les enfants du groupe T2 sont âgés de 11 à 12 ans, ce qui correspond approximativement à l'âge chronologique des sujets Dyslexiques. Au moment des enregistrements, les enfants du groupe Dyslexique sont pensionnaires au centre de rééducation « Les Lavandes » à Orpierre (Sisteron), mais sont, pour la plupart, originaires de la région des Bouches du Rhône . Tous les sujets Dyslexiques ont été sélectionnés par les spécialistes du centre à l'aide de différents exercices neuropsychologiques : répétition de mots difficiles, épreuve de verlan portant sur des pseudo-mots, épreuve de jugement de rimes, épreuve de suppression du premier son de mots, épreuve de segmentation phonémique, transcription de pseudo-mots de structure phonémique simple et complexe. Aucun des dix sujets ne présentait à l'examen neuropsychologique, de trouble déficitaire de l'attention. Ces exercices ont ainsi permis d'obtenir un groupe, le plus homogène possible, d'enfants présentant tous une dyslexie développementale phonologique. Les sujets des groupes T1 et T2 ont été enregistrés à Bruxelles (Belgique). Les dix meilleurs lecteurs de chaque niveau de lecture ont été sélectionnés par leur enseignante.

2.2 Corpus

Le corpus utilisé pour l'ensemble de cette étude est composé de phrases (tâche « phrases ») et d'un texte (tâche « texte »), qui ont été construits selon différents critères : l'ensemble du corpus est ludique et d'un niveau de difficulté adapté à des enfants de première année de lecture. D'après les descriptions d'erreurs considérées comme caractéristiques du trouble [Noe76], les enfants dyslexiques présentent des difficultés dans la distinction du voisement, des lieux et des modes d'articulation, des difficultés à distinguer l'ordre de succession des lettres, des difficultés de décodage des graphèmes complexes. Il a paru intéressant de confronter ces enfants à ces difficultés déjà associées au trouble, afin d'essayer de mieux les comprendre et les décrire à partir d'une analyse des paramètres acoustiques et aérodynamiques de leur production de parole.

Les phrases sont au nombre de six, présentent toutes la même structure [CVCV di CVCV ɑ̃kɔʁ]; les consonnes sont des occlusives bilabiales soit sourdes [p] soit sonores [b] les voyelles sont [a], [i] ou [u].

Le texte comporte sur le plan phonétique, la plupart des consonnes et voyelles du français ; il contient en outre des structures syllabiques simples de type CVC (« école ») et complexes de type CCV (« février »), des graphèmes complexes dont la conversion en phonème nécessite un recours à des règles contextuelles précises (« déguisée », « enfant »).

2.3 *Matériel*

La station de travail EVA (Evaluation Vocale Assistée), utilisée pour le recueil des données, permet l'enregistrement simultané de données acoustiques et aérodynamiques ; elle est composée d'un PC, de différents capteurs et instruments de mesure. Le programme destiné à l'étude aérodynamique de la parole offre la possibilité d'étudier les corrélations entre le signal acoustique et les variations aérodynamiques, ce qui permet de faire des inférences sur les gestes articulatoires.

Les débits d'air inspiré et expiré aux lèvres (débit d'air buccal ou DAB) et aux narines (débit d'air nasal ou DAN) ont été recueillis par l'intermédiaire d'un masque et d'embouts en silicone, placés respectivement à l'entrée du conduit vocal et des narines. La pression intra-orale (PIO) a été enregistrée par l'intermédiaire d'une sonde buccale d'un diamètre extérieur de 3.5mm.

La résolution est de 12 bits pour les données acoustiques et aérodynamiques, leur fréquence d'échantillonnage est respectivement de 6250 Hz et 1560 Hz.

Les différents paramètres à enregistrer (signal, débits, pression) ont été calibrés. Les masques, sondes et embouts utilisés lors de l'enregistrement des données aérodynamiques ont été stérilisés.

2.4 *Protocole expérimental*

Les sujets dyslexiques ont été enregistrés en lecture (DL) puis en répétition (DR). Suite à des contraintes d'ordre temporel, les sujets contrôles ont été enregistrés en lecture uniquement. Les enregistrements ont duré environ ¼ d'heure par sujet.

Pour le test de lecture, l'ensemble du corpus a été présenté, phrase après phrase sur des feuilles A4 tenues face à l'enfant à hauteur d'yeux par l'expérimentateur dans un souci d'adaptation aux difficultés potentielles des différents groupes.

En ce qui concerne le test de répétition, le corpus avait été enregistré au préalable sur une cassette audio par un locuteur de sexe masculin, âgé de 57 ans, d'origine méridionale ; un intervalle d'environ 10 s a été laissé entre chaque phrase et la suivante afin de permettre la répétition.

Des contraintes techniques n'ont pas permis un recueil simultané de tous les paramètres aérodynamiques. Ainsi, la tâche « phrases » a été utilisée pour le recueil du signal acoustique, du DAB et de la PIO, tandis que la tâche « texte » a servi de base d'enregistrement du signal acoustique, du DAB et du DAN.

Lors du recueil des données, au début de chaque enregistrement, des explications sur le déroulement de la séance ainsi que sur le fonctionnement de la station ont été données à chacun des enfants. Pendant la calibration des paramètres, ils ont également pu se familiariser avec le masque et les embouts aux narines ainsi qu'avec la sonde buccale. Enfin, pendant l'enregistrement, nous avons veillé à ce que les embouts utilisés avec la station EVA restent bien en place afin d'obtenir les meilleurs tracés possibles.

2.5 *Traitement des données*

Une première transcription phonétique (API) de l'ensemble des données a été opérée. Puis, les données acoustiques et aérodynamiques ont été analysées à partir du logiciel Phonedit. Les différents segments phonétiques ont été identifiés à partir des critères de segmentation habituellement utilisés [Cal89]. Une étiquette a ensuite été attribuée à chacun des segments identifiés. Enfin, suite à l'identification des différents segments, des précisions ont pu être apportées à la transcription initiale, précisions qui ont permis de noter le caractère subtil des particularités observées ; afin de rendre compte de ces particularités, une signification particulière a été attribuée à certains symboles ; ainsi, [ɸ], [β], [ɹ], ont été utilisés pour rendre compte plus de l'absence d'occlusion complète (présence de fuite d'air) que pour transcrire des fricatives ou approximantes canoniques, tandis que [□] reflète la présence de segments proches des semi-voyelles mais pour lesquels nous n'avons pu déterminer avec précision les caractéristiques articulatoires. C'est cette seconde transcription qui a été utilisée pour l'élaboration d'un tableau de production pour chacun des trois groupes et par type de tâche.

Plusieurs analyses ont été effectuées : au cours de l'analyse des erreurs de production, la transcription des productions de chacun des différents sujets a été comparée avec une transcription prototypique du corpus. Au cours de cette analyse, seules ont été étudiées les erreurs de type phonologique (et non les erreurs de type sémantique ou de conversion). Un segment a été considéré comme erroné lorsque sa réalisation ne correspondait pas au prototype ou lorsqu'il avait été omis ou déplacé. Un tableau d'erreurs par groupe et par type de tâche a ensuite été constitué, où apparaissent le segment prototype, les différentes réalisations et leur nombre pour chaque sujet.

	non continu		continu	
	voise	non voise	voise	non voise
	Sujets dyslexiques : réalisations en lecture			
[b]	56,70%	8,30%	26,70%	
	Sujets dyslexiques : réalisations en répétition			
[b]	53,30%	5,80%	36,70%	
	Sujets témoins 1 : réalisations en lecture n			
[b]	61,10%	24,10%	10,20%	
	Sujets témoins 2 : réalisations en lecture			
[b]	88,30%	0,80%	10%	

Enfin différentes mesures ont été effectuées, uniquement à partir de la tâche « phrases » pour les trois groupes ; des contraintes méthodologiques ont rendu inexploitables les données recueillies à partir de la tâche « texte ». Les mesures effectuées à partir de la tâche « phrases » ont ainsi constitué la base des analyses des paramètres acoustiques et aérodynamiques du corpus ; elles concernent la durée : des voyelles en position préconsonantique (V/-C), de la tenue des consonnes intervocaliques (C/V-V), du VOT sur [p] et [k], ainsi que les maxima de PIO sur [p] et [b], et les maxima de DAB sur tous les segments vocaliques et consonantiques.

3 RESULTATS

3.1 Analyse des erreurs de production

Cette étude a concerné la fréquence d'apparition des types d'erreurs les plus largement représentés dans les trois groupes (plus de 10%) pour les tâches « phrases » et « texte ».

Au sein du groupe Dyslexique, pour la tâche « phrases », on observe une quantité d'erreurs nettement plus importante en répétition qu'en lecture en ce qui concerne les segments [p] (54.2% et 26.7%) et [b] (46.7% et 43.3%). Des erreurs sur les segments vocaliques ont été observées en lecture uniquement. L'apico-dentale sonore [d] est réalisée [ɹ] quand cette consonne est placée avant la voyelle [i] (d/-i) : 46,7% en répétition, 47,5% en lecture.

Si l'on compare ensuite les groupes Dyslexique, T1 et T2, le pourcentage de réalisations correctes des phonèmes cibles est en accord avec le niveau de lecture des différents groupes. Cependant, si l'on observe pour tous les enfants, des erreurs de voisement asymétriques (les voisées sont réalisées non-voisées et non l'inverse), elles apparaissent en nombre moins important dans le groupe Dyslexique qui montre une tendance nette à produire des fricatives bilabiales sonores [β] ou des réalisations [□] pour l' occlusive bilabiale sonore [b] (11.7% et 15%). De même, la réalisation [ɹ] pour le segment [d] est nettement plus fréquente chez les sujets dyslexiques que chez les sujets T1 et T2 : 47.5% contre des pourcentages inférieurs à 10%.

La tâche « texte » permet d'observer des erreurs de voisement pour les trois groupes, des omissions pour les groupes T1 et T2, des occlusions incomplètes pour le groupe Dyslexique.

3.2 Analyse des données acoustiques

Cette analyse, en cours à l'heure actuelle, n'a concerné jusqu'à présent que les segments identifiés comme conformes au modèle ; les résultats concernant les segments sur lesquels portent les erreurs pourront être présentés lors de la conférence.

Réalisée à partir de la tâche « phrases » pour les segments correctement produits, cette étude révèle dans un premier temps que les durées de la tenue des consonnes intervocaliques [p] et [b] (C/V-V) sont de façon générale plus importantes en lecture qu'en répétition chez les sujets du groupe Dyslexique. Les durées moyennes du VOT, sont, elles, équivalentes dans les deux conditions. De même, aucune différence entre les deux conditions de production n'est à noter en ce qui concerne les durées des segments vocaliques V/-C.

Une comparaison des trois groupes Dyslexique, T1 et T2 permet d'observer que les durées de la tenue des segments [p], [b] et [d] en position intervocalique sont plus importantes pour les sujets du groupe T1, puis pour les sujets du groupe Dyslexique, enfin ceux du groupe T2 présentent les durées les plus brèves. En revanche, le groupe Dyslexique présente les durées les plus brèves pour le segment [k]. Les durées moyennes du Vot sont plus longues en lecture pour les sujets du groupe Dyslexique que pour les sujets des groupes T1 et T2. En ce qui concerne les durées des voyelles en position intervocalique V/-C, on peut remarquer une différence très nette entre les trois groupes en lecture : les sujets du groupe T1 se distinguent par des durées importantes alors que les sujets du groupe T2 se caractérisent par des durées brèves, les sujets du groupe D étant en position intermédiaire.

3.2 Analyse des données aérodynamiques

Les mesures de DAB ainsi que les mesures de Pio sont équivalentes chez les sujets du groupe Dyslexique quelle que soit la condition de production.

En revanche, la comparaison des résultats des groupes D, T1 et T2 en lecture permet de noter que ces mêmes sujets présentent par rapport aux sujets des deux groupes contrôles, des valeurs de PIO significativement moins élevées pour les segments [p] et [b].

La figure 1 illustre les particularités acoustiques et articulatoires observées chez un enfant dyslexique, en lecture de phrase, pour la séquence [padipa].

4 DISCUSSION

La dyslexie depuis sa première description en 1896, est définie comme un trouble spécifique et durable de l'acquisition du langage écrit. Pourtant on trouve dans la littérature plusieurs travaux qui mettent ce déficit en relation avec le langage oral. Ceux qui ont essentiellement

porté sur la perception de la parole ont mis en évidence un trouble subtil de la perception chez le dyslexique qui pourrait être responsable du déficit de conscience phonologique [Mor91]. Les quelques travaux menés en production de la parole [Hei96] ont pour leur part conduit à penser qu'un trouble subtil pouvait entraîner une mauvaise mise en place des processus du langage oral, ce qui conduirait à des difficultés sévères et durables dans l'acquisition de la lecture. Enfin, les travaux d'imagerie cérébrale [Pau96] ont mis en évidence le rôle de l'aire de Broca (motricité) dans des tâches silencieuses de lecture ou de manipulation phonémique, ce qui permet d'affirmer l'implication des systèmes de production dans le déficit.

L'analyse multiparamétrique des productions des sujets dyslexiques apporte des arguments supplémentaires en faveur de ces différentes données : on pourrait situer l'origine du déficit au moment de l'acquisition du langage oral puisque les erreurs de production sont plus nombreuses en répétition qu'en lecture. Ce déficit, auparavant décrit par le biais d'erreurs de type catégoriel [Noe76], est observable au niveau infra phonémique, ce qui révèle peut-être un déficit au niveau de la réalisation articulatoire des segments, plus qu'au niveau de leur encodage : les erreurs concernent en effet la structure interne des consonnes, touchant de façon inter dépendante le mode ou le voisement. Par exemple le segment [b] est réalisé [p] (voisement) ou [ß] (absence d'occlusion) on a donc soit une absence de voisement qui va de pair avec une occlusion correctement réalisée, soit le voisement est conservé alors que l'occlusion est incomplète. Ce phénomène est d'un point de vue articulatoire explicable par des contraintes aérodynamiques : il est difficile de maintenir un différentiel de pression, entre la pression sous glottique (PSG) et la pression intra orale (PIO), nécessaire à la vibration des cordes vocales alors que l'occlusion a pour effet d'équilibrer ces deux pressions de part et d'autre de la glotte. De plus, les valeurs de PIO des segments [p] et [b], même si il s'agit de mesures effectuées sur des segments correctement réalisés, viennent confirmer cette difficulté articulatoire puisqu'elles suggèrent l'existence d'une fuite potentielle.

Cette étude préliminaire des productions articulatoires des enfants atteints de dyslexie développementale phonologique apporte ainsi des arguments supplémentaires en faveur de l'idée que l'origine du déficit en lecture se situerait au moment de l'acquisition du langage oral. En revanche cette étude ne permet pas de déterminer à quel niveau se situe ce déficit : s'agit-il d'une difficulté dans la programmation articulatoire, dans la coordination des gestes ou encore dans la boucle sensori-motrice (feed-back). Si ces données montrent bien l'implication, d'un point de vue causal, des systèmes de production dans la genèse du trouble, d'autres analyses articulatoires complétées par une étude en perception par exemple, ou encore la mise en relation des particularités articulatoires observées avec l'intensité, le degré de gravité du trouble pourraient permettre de mieux comprendre leur implication causale éventuelle.

BIBLIOGRAPHIE

[Ale91] Alexander A.W. Andersen H.G. Heilman P.C. Voeller K Torgesen J.K. (1991), Phonological awareness training and remediations of analytic decoding deficits in a group of severe dyslexics, Ann. Dyslexia,41, pp. 193-206.

[Cal89] Calliope (1989), La parole et son traitement automatique, Masson

[hab97] Habib M. (1997), Dyslexie: le cerveau singulier, coll. Neuropsychologie, Marseille, Solal.

[Hei96] Heilman K.M. Voeller K. & Alexander A.W. (1996), Developmental dyslexia: a motor articulatory feed-back hypothesis, Ann. Neurologique, 39, pp. 407-412.

[Lec85] Lecocq P. (1985), Apprentissage de la lecture et dyslexie, Ed. Mardaga, Paris.

[Mo81] Montgomery D. (1981), Do dyslexics have difficulty accessing articulatory information, Psychologie Res., 43, pp. 235-243.

[Mor91] Morais J. (1991), Constraint on the development of phonemic awareness, In Brady S.A. & Shankweiler D.P. (Eds), Phonological processes in literacy, A tributez to Isabelle Y. Liberman, Lawrence Erlbaum Associates, New Jersey, pp. 67-83.

[Noe76] Noël J.M. (1976), La dyslexie en pratique éducative, Ed. Doin, Paris.

[Pau96] Paulesu E. & al. (1996), Is developmental dyslexia a disconnection syndrome? Evidence from PET scanning Language comprehension in language learning impaired children improved, Brain, 119, pp.143-157.

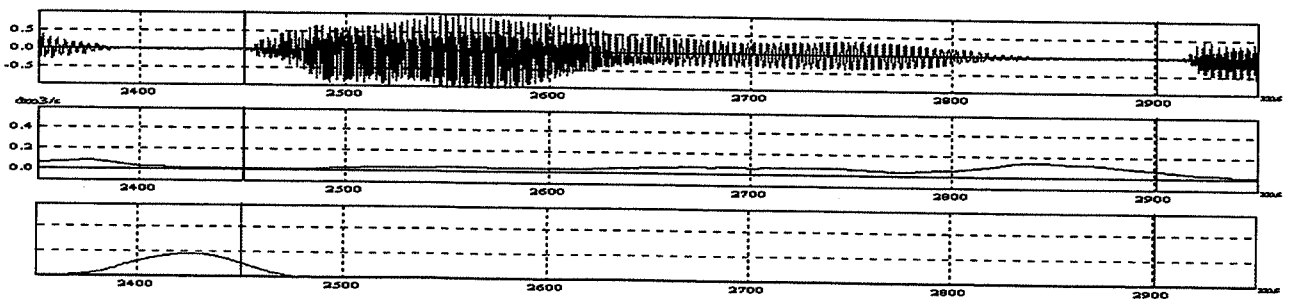


Figure 1 : signal acoustique, DAB et PIO pour la séquence [padipa] par un sujet dyslexique en lecture tâche « phrases ».

Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire

Slim OUNI et Yves LAPRIE

LORIA - UMR 7503
BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex - France
Mél: ouni,laprie@loria.fr

ABSTRACT

This paper presents an articulatory codebook construction method which gives rise to a good articulatory space coverage with a limited number of points. It is a new representation of the articulatory space by hypercubes. A hypercube is a region represented by a few number of points. We also present an interpolation method to retrieve points from a hypercube and an inversion method based on the same interpolation method. The advantage of the codebook and the inversion method is its strength against articulatory-to-acoustic mapping non-linearity problems.

1 INTRODUCTION

L'inversion acoustico-articulatoire consiste à récupérer les paramètres articulatoires décrivant la forme du conduit vocal à partir du signal de parole. L'une des familles de méthodes d'inversion est celles des méthodes par tabulation qui exploitent un dictionnaire de formes articulatoires indexées par les paramètres acoustiques. Ce dictionnaire est construit en utilisant un synthétiseur articulatoire qui à partir des paramètres articulatoires calcule les paramètres acoustiques.

Notre méthode d'inversion repose sur le modèle articulatoire de Maeda [MAE79] et consiste à régulariser les trajectoires articulatoires initiales obtenues à l'aide d'une méthode par tabulation. Le dictionnaire de formes nous permet de construire les trajectoires initiales, et conditionne donc fortement la qualité finale de l'inversion. Il faut donc construire le dictionnaire de formes articulatoires avec une grande attention. Pour ce faire, plusieurs méthodes existent:

- échantillonnage aléatoire des paramètres articulatoires [SCH90] ;
- échantillonnage autour des trajectoires liant les formes de base correspondant aux voyelles [LAR88] ;
- échantillonnage implicite de l'espace articulatoire pour entraîner un système neuromimétique comme le modèle "forward" [LAB95].

Pour la construction de notre dictionnaire, nous utilisons un système d'hypercubes. Comme nous le verrons, notre méthode est destinée à réduire au minimum l'espace et le temps nécessaires à une exploration systématique de l'espace articulatoire. Le dictionnaire final est certes encore volumineux, mais permet d'être assuré que l'échantillonnage articulatoire n'a plus d'influence sur l'inversion, ce qui est loin d'être le cas pour les autres méthodes. A titre d'exemple nous montrons dans notre étude expérimentale la comparaison entre la trajectoire récupérée avec le dictionnaire hypercubique et celle récupérée avec un dictionnaire de 600000 formes choisies aléatoirement. En effet, une trajectoire linéaire dans l'espace articulatoire ne s'accompagne pas forcément d'une trajectoire linéaire dans l'espace acoustique [OUN99]. Par conséquent, une région qui présente ce genre de non-linéarité peut être omise s'il n'y a pas suffisament de points couvrant cette région. En fait, si nous voulions faire un échantillonnage régulier et fin des sept paramètres du modèle articulatoire de Maeda dans l'intervalle $[-3\sigma, 3\sigma]$ (σ étant l'écart type) avec un pas d'échantillonnage relativement grossier de 1/3 σ, nous obtiendrions $(19^7) \equiv 900$ millions de points, ce qui est très coûteux pour les machines actuelles en espace de stockage comme en temps d'accès.

L'idée est d'avoir un échantillonnage moins coûteux, mais précis. Il doit être précis, pour pouvoir être sûr de récupérer toutes les solutions possibles, pour étudier l'influence articulatoire des contraintes ajoutées à l'inversion et pour pouvoir trouver la trajectoire articulatoire qui est à l'origine du signal de la parole à inverser. En effet, les méthodes d'inversion existantes exploitent, voire abusent, de l'effet compensatoire du conduit vocal ce qui peut fausser l'interprétation des résultats. Dans les paragraphes qui suivent nous présentons notre méthode de construction du dictionnaire avec une meilleure couverture de l'espace articulatoire et ensuite les outils d'inversion pour un tel dictionnaire.

1. LA CONSTRUCTION DU DICTIONNAIRE DE FORMES

1.1 Le dictionnaire hypercubique

L'idée s'inspire du fait que la relation articulatoire-acoustique (qu'on notera $\mathcal{M}$) est non-linéaire. C'est un problème qui doit être pris en compte si nous voulons obtenir une couverture efficace de l'espace articulatoire. Nous rappelons que la non-linéarité de la relation $\mathcal{M}$ est inévitable vu qu'elle est liée à la nature physique et géométrique du conduit vocal [CHA84]. Pour cela, nous décomposons l'espace articulatoire d'une manière fine dans les régions où la relation $\mathcal{M}$ est fortement non-linéaire. Dans ce but, nous utilisons les hypercubes. Un hypercube d'ordre N est une région d'un espace de dimension N délimitée par des hyperplans. L'espace articulatoire sera représenté par une arborescence d'hypercubes. Chaque hypercube représente une région de l'espace articulatoire où la relation $\mathcal{M}$ peut être considérée comme linéaire. Le lecteur peut trouver dans [OUN99] les détails de la construction de l'hypercube. Rappelons ici seulement le principe de la méthode.

1.2 Le principe de la méthode

Nous supposons que tout l'espace articulatoire est contenu dans un hypercube. Si la relation $\mathcal{M}$ est non-linéaire à l'intérieur de cet hypercube, ce dernier est décomposé en sous-hypercubes. Pour chaque sous-hypercube, nous testons de nouveau la linéarité. Si la relation est quasi-linéaire, nous gardons cet hypercube sinon, nous le décomposons à nouveau. Cette procédure est répétée récursivement jusqu'à l'obtention d'un hypercube de taille suffisamment petite pour pouvoir considérer que le comportement de la relation $\mathcal{M}$ dans cet hypercube est linéaire.

1.3 *Le test de linéarité*

Le test proposé par Charpentier [CHA84] consiste à calculer la courbure acoustique le long d'un chemin articulatoire à l'intérieur de la région à explorer. Cette méthode acceptable dans le cas d'un modèle de fonction d'aire qui utilise peu de paramètres conduirait à des calculs trop longs dans notre cas. C'est pourquoi nous utilisons le test suivant. Pour tous les segments qui relient les sommets d'un hypercube, nous considérons les milieux de ces segments et nous interpolons linéairement les valeurs acoustiques correspondantes. Ensuite, nous comparons ces valeurs avec celles calculées directement avec le synthétiseur articulatoire. Si la différence entre les valeurs acoustiques synthétisées et les valeurs acoustiques interpolées est inférieure à un seuil prédéfini $\Delta\varepsilon$, la relation $\mathcal{M}$ dans cet hypercube est considérée comme linéaire. Nous disons que $\mathcal{M}$ est linéaire avec une marge d'erreur de $\Delta\varepsilon$ dans le domaine acoustique. Pour un hypercube de dimension 7, nous avons 128 sommets (2^7) et le nombre de segments possibles entre ces sommets est 8128, ce qui correspond au nombre de tests. Nous supposons que ce test de linéarité est suffisant.

1.4 *La description du dictionnaire hypercubique.*

En résumé, un hypercube est défini par ses sommets qui sont des vecteurs de l'espace articulatoire. Dans un dictionnaire hypercubique, un hypercube est représenté par un sommet origine, la longueur d'un des cotés (avec ces deux informations seulement, nous pouvons construire l'hypercube) et les valeurs acoustiques des sommets. En conséquence, le dictionnaire est composé par des hypercubes plus ou moins fins, selon la linéarité de la relation $\mathcal{M}$ dans la région représentée par l'hypercube. Plus l'hypercube est grand, plus la relation $\mathcal{M}$ est linéaire.

1.5 *L'interpolation dans un hypercube*

Nous pouvons récupérer toutes les informations dont nous avons besoin à partir des sommets de l'hypercube. En effet, la seule connaissance des vecteurs articulatoires et des paramètres acoustiques leurs correspondant, nous permet de retrouver toutes les informations concernant les vecteurs articulatoires se trouvant à l'intérieur de cet hypercube par une interpolation à partir des sommets. Pour rendre l'interpolation plus robuste et plus précise, nous interpolons par rapport au sommet le plus proche du vecteur dont nous cherchons les paramètres acoustiques en calculant le gradient en ce sommet. Considérer le sommet le plus proche permet, en effet, de renforcer l'hypothèse de linéarité. Le sommet le plus proche est retrouvé à partir des 128 sommets de l'hypercube.

Soit $\vec{P}$ le vecteur articulatoire (ses composantes sont les paramètres articulatoires du modèle de Maeda) dont nous cherchons son correspondant acoustique $\vec{F}$ (ses composantes sont les trois premiers formants) par interpolation dans l'hypercube H_c.

$$\vec{F}=\mathcal{M}(\vec{P}) \qquad (1)$$

L'interpolation au sens du gradient par rapport au sommet le plus proche P_0 est donnée par l'équation suivante :

$$\vec{F} = \vec{F}_0 + \nabla\vec{F}\cdot(\vec{P}-\vec{P}_0) \qquad (2)$$

Où $\vec{F}_0 = \mathcal{M}(\vec{P}_0)$ et $\nabla\vec{F}$ est le gradient de $\vec{F}$.

Connaissant $\vec{P}$, P_0 et F_0, nous calculons $\vec{F}$. Grâce à l'équation (2), pour tout vecteur articulatoire se trouvant à l'interieur de l'hypercube, nous pouvons retrouver toute l'information acoustique qui lui est relative par interpolation.

1.6 *Vérification expérimentale de l'interpolation*

Afin de tester cette méthode d'interpolation, nous avons généré une trajectoire acoustique qui a été synthétisée avec le synthétiseur articulatoire à partir d'une trajectoire articulatoire. Par ailleurs, cette trajectoire est interpolée à partir du dictionnaire hypercubique. Nous comparons la proximité du signal acoustique correspondant à la trajectoire synthétisée et le signal correspondant à la trajectoire interpolée.

Nous obtenons de bons résultats du point de vue de la proximité acoustique, comme le montre la figure 1. En effet, pour la construction du dictionnaire nous avons fixé une marge d'erreur assez grande pour le test de linéarité (50Hz pour le premier formant, 75Hz pour le deuxième formant et 100Hz pour le troisième formant). Nous avons généré 37 trajectoires articulatoires. L'erreur moyenne ne dépasse pas 10Hz pour les deux premiers formants et 20Hz pour le troisième formant. Ceci constitue une bonne approximation formantique et est rassurant à propos de la qualité de l'interpolation.

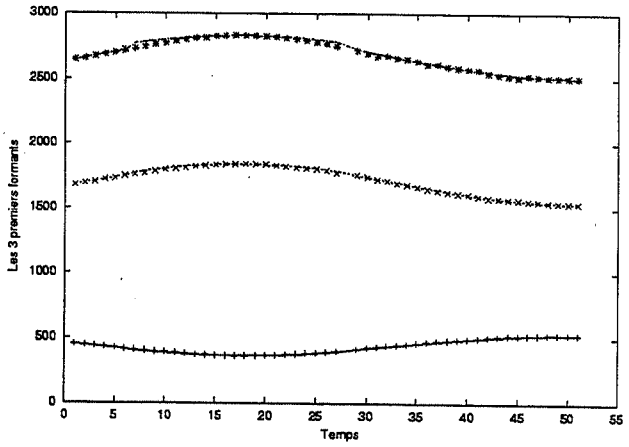


Figure 1 - Le paramètre correspondant à la mâchoire varie linéairement (les autres paramètres restent constants). Nous représentons la trajectoire synthétisée (trait fin) et interpolée (points) dans l'espace acoustique des trois premiers formants.

2 UTILISATION DU DICTIONNAIRE HYPERCUBIQUE POUR L'INVERSION

Etant donné un signal acoustique, nous voulons maintenant récupérer les paramètres articulatoires qui sont à l'origine de ce signal. Les solutions obtenues doivent vérifier les deux critères suivants:

- une bonne proximité acoustique avec les données de départ;
- la régularité des trajectoires articulatoires. Si cela est possible, nous voulons retrouver les trajectoires articulatoires qui sont à l'origine du signal acoustique mesuré ou, au moins, une solution proche de l'originale.

Nous rappelons qu'un hypercube représente une région où la relation $\mathcal{M}$ est quasi-linéaire. L'image d'un hypercube articulatoire est donc un polygone inscrit dans un hyperplan acoustique.

Le signal est décomposé en segments de quelques millisecondes chacun. Chaque segment constitue une entrée acoustique à inverser. Nous cherchons les hypercubes du dictionnaire

hypercubique dont l'image par $\mathcal{M}$ (donc dans l'hyperplan acoustique) contient cette entrée acoustique. Nous augmentons la taille des hyperplans acoustiques de quelques Hertz afin d'éviter les problèmes aux limites. A partir du dictionnaire de formes nous récupérons tous les hypercubes répondant à cette requête. En effet, une entrée acoustique peut appartenir à plusieurs hypercubes. Mais, chaque hypercube ne fournit qu'une seule solution (hypothèse de linéarité dans un hypercube).

Soit $\vec{F}$ le vecteur acoustique (représenté par les trois premiers formants) à inverser. Soit H_c un hypercube tel que $\vec{F} \in \mathcal{M}(H_c)$. Soit $\vec{P}$ le vecteur articulatoire (représenté par les sept paramètres du modèle articulatoire de Maeda) cherché associé à $\vec{F}$.

Nous utilisons le même principe pour l'inversion que celui de l'interpolation. Nous faisons l'hypothèse que le point inversé est proche d'un sommet P_0. Ce dernier est choisi comme étant le sommet ayant son vecteur acoustique F_0 le plus proche de $\vec{F}$. L'inversion en utilisant l'interpolation au sens du gradient se fait en résolvant l'équation suivante (qui découle de l'équation (2)):

$$\vec{F} - \vec{F}_0 = \nabla\vec{F} \cdot (\vec{P} - \vec{P}_0) \qquad (3)$$

L'équation (3) est un système d'équations linéaires qui peut être écrit de la manière suivante:

$$\begin{pmatrix} F^1 - F_0^1 \\ F^2 - F_0^2 \\ F^3 - F_0^3 \end{pmatrix} = \begin{pmatrix} \frac{\partial F^1}{\partial \alpha_1} & \frac{\partial F^1}{\partial \alpha_2} & \cdots & \frac{\partial F^1}{\partial \alpha_1} \\ \frac{\partial F^2}{\partial \alpha_1} & \frac{\partial F^2}{\partial \alpha_2} & \cdots & \frac{\partial F^2}{\partial \alpha_7} \\ \frac{\partial F^3}{\partial \alpha_1} & \frac{\partial F^3}{\partial \alpha_2} & \cdots & \frac{\partial F^3}{\partial \alpha_7} \end{pmatrix} \cdot \begin{pmatrix} (P^1 - P_0^1) \\ (P^2 - P_0^2) \\ (P^3 - P_0^3) \\ (P^4 - P_0^4) \\ (P^5 - P_0^5) \\ (P^6 - P_0^6) \\ (P^7 - P_0^7) \end{pmatrix}$$

Où F^i, F_0^i représentent les ièmes composantes des vecteurs $\vec{F}$ et F_0 et P^i, P_0^i sont les ièmes composantes des vecteurs $\vec{P}$ et P_0

Pour résoudre un tel système, dont le nombre d'équations est inférieur au nombre des inconnues, nous avons eu recours à un alogrithme se basant sur la méthode SVD (décomposition en valeurs singulières) [GOL89]. Cette méthode permet d'obtenir toutes les solutions du système d'équations. Dans notre cas, nous voulons une solution dans le voisinage immédiat du sommet P_0 pour respecter l'hypothèse de calcul du jacobien qui est la dérivée calculée au sommet P_0. Nous choisissons donc le point de l'espace solution le plus proche de $\vec{P}$, c'est à dire le point de l'hypercube pour lequel la distance $(\vec{P} - P_0)$ est minimale[1]. Grâce à cet algorithme nous obtenons le point inverse. A ce stade nous vérifions l'hypothèse de départ concernant la proximité du vecteur inverse $\vec{P}$ par rapport au sommet P_0. Pratiquement il faut donc tester l'hypothèse de proximité pour tous les sommets (2^7=128 sommets) ou s'arrêter dès que cette hypothèse est bien vérifiée. Ce processus est appliqué à tous les hypercubes H_i tels que $\vec{F} \in \mathcal{M}(H_i)$, et de même, il est appliqué à l'ensemble des entrées acoustiques.

[1] Plus précisément, SVD permet d'obtenir l'espace des solutions sous la forme d'un point Q et des vecteurs de base qui correspondent au noyau du système. Le point Q est le point de l'espace solution le plus proche de l'origine [GOL89] qui est dans notre cas le sommet de l'hypercube considéré P_0. Comme nous nous sommes placés dans l'hypothèse de trouver la solution la plus proche du sommet P_0, Q est donc le point cherché à condition qu'il appartienne à l'hypercube.

2.1 *Contraindre l'inversion*

Pour que l'inversion réussisse, trois conditions doivent être vérifiées:

- $F_0 = \mathcal{M}(P_0)$ est l'image acoustique du sommet P_0, la plus proche de $\vec{F}$ (vecteur acoustique à inverser): $r_0 = \min d(r, r_i)$, F_i étant les paramètres acoustiques correspondant au sommet i (i=1..128),
- P_0 reste toujours le sommet le plus proche de $\vec{P}$ (vecteur articulatoire), résultat de l'inversion qui est calculé par interpolation: $P_0 = \min d(P, P_i)$, avec P_i le sommet i de l'hypercube (i=1.. 128)
- $\vec{P}$, le vecteur articulatoire trouvé par inversion, doit être à l'intérieur de l'hypercube qui a servi pour l'inversion.

La première condition est la définition même du sommet le plus proche. La deuxième condition traite le cas où la relation $\mathcal{M}$ n'est pas suffisament linéaire dans H_c (à cause des erreurs dans le test de linéarité). Ces deux contraintes sont concurrentes: elles assurent la proximité du sommet le plus proche dans les deux espaces à la fois. Ce qui améliore la linéarité de la relation $\mathcal{M}$ dans l'hypercube. La troisième condition élimine les erreurs dues au fait que l'inversion donne des résultats en dehors de l'hypercube articulatoire considéré. Si l'une des trois premières conditions n'est pas vérifiée, nous rejettons la solution. En combinant l'interpolation par rapport au sommet le plus proche au sens du gradient et ces trois conditions, nous augmentons la précision de l'inversion.

3 ETUDE EXPÉRIMENTALE ET DISCUSSION

Afin de tester la qualité de l'inversion, nous avons construit un dictionnaire de formes de dimensions réduites avec seulement les 5 premiers paramètres. Cela ne change en rien notre méthode d'inversion, car seul le dictionnaire change et faire les tests avec un dictionnaire de dimensions 5 permet d'examiner les solutions plus facilement. Ce dictionnaire contient 5526 hypercubes de dimensions 5. Le nombre de sommets total est donc 176.832. Nous avons pris comme marge d'erreur pour le test de linéarité le triplet (50Hz,75Hz,100Hz) pour les trois premiers formants. Des trajectoires tests ont été générées avec le synthétiseur articulatoire en faisant varier un ou plusieurs paramètres articulatoires. Nous utilisons des trajectoires simulées pour pouvoir faire une comparaison pertinente puisque l'on compare les résultats à une trajectoire connue. Nous passons à notre procédure d'inversion le signal acoustique synthétisé, et obtenons en sortie toutes les solutions obtenues à partir du dictionnaire hypercubique. Nous vérifions deux critères:

- la proximité acoustique par rapport au signal à inverser.
- le fait que la trajectoire articulatoire initiale soit parmi les solutions obtenues ou non.

Pour évaluer la proximité acoustique, nous générons le signal acoustique à partir des résultats de l'inversion et nous le comparons au signal acoustique original. Le deuxième critère est destiné à assurer que nous récupérons bien toutes les solutions, et donc nous avons une bonne couverture de l'espace articulatoire. Si l'inversion fonctionne correctement nous devons retrouver parmi ces solutions la trajectoire originale.

Dans la figure 2, nous présentons les solutions de l'inversion d'un signal acoustique obtenu en faisant varier sinusoïdalement le paramètre articulatoire de la mâchoire et en laissant les autres constants. Pour évaluer acoustiquement l'inversion nous avons resynthétisé le signal acoustique à partir des paramètres articulatoires obtenus par inversion. Pour chaque trame de parole, nous obtenons plusieurs solutions proches de la

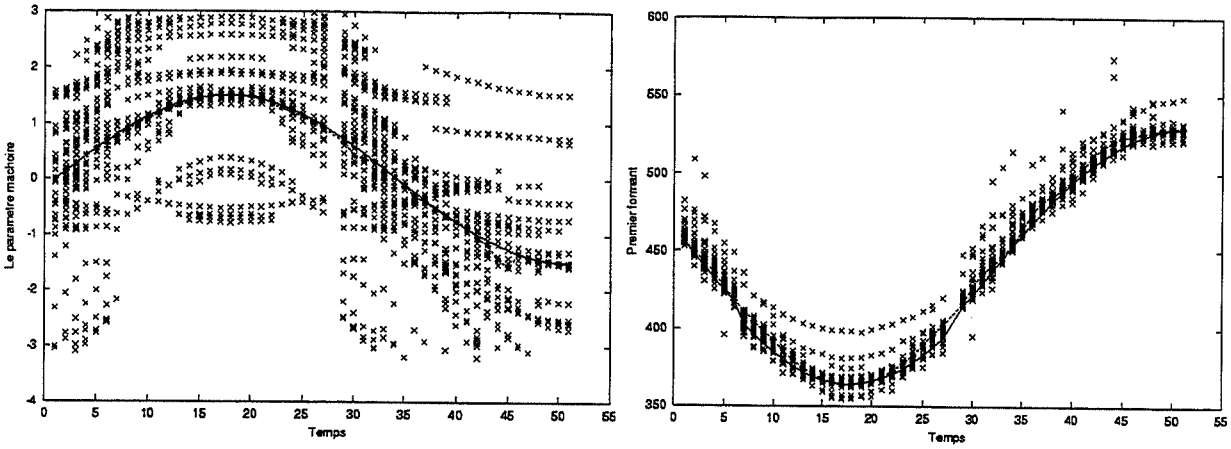


Figure 2 - Représentation des solutions de l'inversion dans l'espace articulatoire (le 1er graphique) et dans l'espace acoustique, premier formant (le 2ème graphique). Les trajectoires initiales sont représentées en trait continu, et les solutions par des points (x).

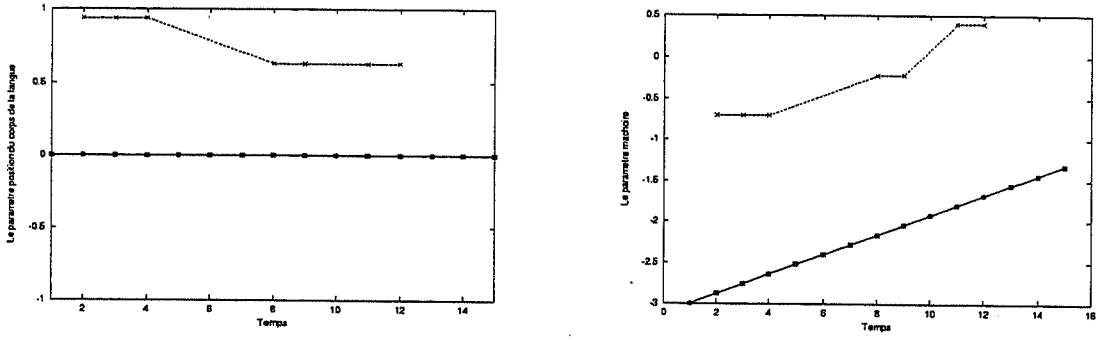


Figure 3 - Comparaison entre l'inversion par un dictionnaire à échantillonnage aléatoire et l'inversion par le dictionnaire hypercubique. Nous présentons la trajectoire dans le plan paramètre position du corps de la langue (1er graphique) et dans le plan du paramètre mâchoire (2ème graphique). La trajectoire de départ (—) et la solution de l'inversion par le dictionnaire hypercubique (▪) se superposent. La solution de l'inversion par le dictionnaire à échantillonnage aléatoire (×) est loin de la trajectoire de départ.

trajectoire acoustique initiale. Dans l'espace articulatoire, nous avons plusieurs solutions possibles. Parmi ces solutions, nous retrouvons bien la trajectoire de départ, qui correspond à la sinusoïde du départ, grâce à une méthode de lissage non-linéaire [LAP98]. Le point fort de cette méthode d'inversion est qu'elle ne contraint pas implicitement le processus d'inversion. Il est donc possible d'étudier très précisement comment l'introduction de contraintes d'origine physiologiques ou acoustiques influence l'inversion de manière à récupérer les trajectoires articulatoires proches des trajectoires réalisées par le locuteur. Pour cela, nous envisageons l'introduction des masses différentes suivant les articulateurs et faire un apprentissage à partir des données réelles .

Nous terminons notre étude expérimentale par une comparaison entre l'inversion avec le dictionnaire hypercubique et l'inversion avec un dictionnaire à échantillonnage aléatoire. Comme pour les autres tests, nous effectuons l'inversion avec les deux dictionnaires de formes, et nous appliquons le même algorithme de lissage pour obtenir la solution. Comme cela apparaît sur la figure 3, pour le paramètre mâchoire et le paramètre position du corps de la langue, la trajectoire obtenue en utilisant le dictionnaire hypercubique coincide avec la trajectoire de départ, alors que la deuxième, obtenue en utilisant le dictionnaire de formes aléatoires de 600000 formes, ne donne pas l'inversion de tous les points et le résultat est même totalement différent de la trajectoire articulatoire initiale. La deuxième solution est obtenue en abusant de l'effet de compensation.

Comme la taille du dictionnaire complet, du moins avec la précision que nous nous sommes imposés, est trop importante pour une machine traditionnelle et que notre algorithme se prête bien au parallélisme, nous sommes en train d'implanter l'algorithme sur une machine parallèle.

BIBLIOGRAPHIE

[CHA 84] Charpentier F. (1984), "Determination of the vocal tract shape from the formants by analysis of the artiuclatory -to-acoustic non-linearities", Speech Com. vol. 3, pp.291-308.

[GOL 89] Golub G. H. & Van Loan C. F. (1989), "Matrix computations", 2ème edition, JHU Press, §8.3, chap. 12.

[LAB 95] Laboissière R. & Galvan A. (1995), "Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences", ICPhS'95, vol. 1, pp. 358-361.

[LAP 98] Laprie Y. & Mathieu B. (1998) "A variational approach for etimating vocal tract shapes from speech signals", ICASSP98, vol. 2, p929-932.

[LAR 88] Larar J. N. & Sondhi M. M. "Vector quantization of the articulatory space", IEEE Trans. Acou., Speech, Signal Processing, vol. 36 n°12, pp. 1812-1818.

[MAE 79] Maeda S. (1979) "Un modèle articulatoire de la langue avec des composantes linéaires", JEP 79, p152-162.

[OUN 99] Ouni S. & Laprie Y. (1999), "Design of hypercube codebooks for the acoustic-to-articulatory inversion respecting the non-linearities of the articulatory-to-acoustic mapping", Eurospeech'99, vol. 1, pp. 141-144.

[SCH 90] Schroeter J., Meyer P. et Parthasarathy S. (1990), "Evaluation of improved articulatory codebooks and codebook access distance measures", ICASSP'90, vol. 1, pp. 393-396.

Quels paramètres peut-on mesurer pour évaluer un modèle de voix pathologique?

S. Hans, J. Vaissière, D. Brasnu

IPLGA, UPRESA-CNRS 7018
Service d'Oto-Rhino-Laryngologie, Hôpital Laënnec, 42, rue de Sèvres, 75007 Paris, France
Mél: sthans@club-internet.fr

ABSTRACT

*Objective:*The aim of this prospective study was to compare the interest of respiratory, acoustic, aerodynamic measurements and videofiberscopy with stroboscopy (VFS) for the evaluation of dysphonic voices. *Subjects and Methods:* Four patients with bilateral vocal fold paralysis and treated by CO2 Laser posterior transverse cordotomy (LPTC) were recorded pre- and postoperatively at 1, 3, 6, 12 and 24 months. Frequency features (F0, jitter, shimmer and HNR), speech duration parameters, laryngeal aerodynamic parameters (intraoral pressure, oral airflow and sound pressure level) and VFS parameters were measured noninvasively. *Results:*Postoperatively at 1 and 3 months, frequency features were indectable by standard commercialized algorithms. In contrast, respiratory, aerodynamic and VFS parameters could always be performed and attested improvement in the follow-up. *Conclusion:*Laryngeal aerodynamic parameters and videofiberscopy with stroboscopy in contrast to acoustic measurements can be used objectively to follow patients longitudinally after LPTC.

1. INTRODUCTION

Notre modèle est la voix produite par les patients atteints de paralysie récurrentielle bilatérale (PRB) traités par cordotomie transverse postérieure au Laser CO2 (CTP). Le traitement de la PRB est un compromis entre les différentes fonctions du larynx : la respiration, la phonation et la déglutition. Evaluer la voix de ces patients a un triple objectif :i) comparer à partir des données objectives les différentes techniques chirurgicales réalisées par voie endoscopique au Laser CO2 (aryténoïdectomie totale ou partielle et CTP), ii) pouvoir guider la rééducation orthophonique à partir de ces données objectives, iii) préserver la qualité de vie du patient. La technique de cordotomie transverse postérieure au Laser CO2 a été introduite par Dennis et Kashima [Den89]. Le but de cette étude prospective était d'analyser l'apport des paramètres respiratoires, acoustiques, aérodynamiques, et morphodynamiques par vidéofibroscopie laryngée avec stroboscopie dans l'évaluation des fonctions laryngées après cordotomie transverse postérieure au Laser CO2 (CTP).

2. MATÉRIEL ET MÉTHODES

2.1 Patients

Deux sujets masculins (âgés de 69 et 77 ans) et deux sujets féminins (âgées de 52 ans et 74 ans) ont été inclus dans cette étude prospective. Ces quatre patients étaient atteints d'une paralysie récurrentielle bilatérale en adduction et ont été traités par CTP bilatérale. Aucune trachéotomie n'a été réalisée et aucune complication postopératoire n'est survenue.

2.2 Méthodologie

Tous les patients ont été enregistrés selon le même protocole non invasif en préopératoire et en postopératoire à 1, 3, 6, 12 et 24 mois.

Analyse respiratoire Les paramètres respiratoires ont été déterminés à partir d'un spiromètre type Gould II. Le volume maximum inspiré pendant la première seconde (VIMS), le volume expiré pendant la première seconde seconde (VEMS), le Peakflow (ou débit de pointe) et la capacité vitale (CV) ont été analysés.

Analyse acoustique Les paramètres retenus pour l'étude de la voix étaient la fréquence fondamentale (F0) et sa déviation standard, le jitter, le shimmer et le rapport signal sur bruit (HNR). La voyelle retenue pour la détermination de ces paramètres fréquentiels était le /a/ réalisé à une intensité et à une fréquence confortable sur une seule expiration. Trois secondes de la portion stable de la voyelle prolongée enregistrée étaient analysées automatiquement par les logiciels MDVP ET CSL de Kay Elemetrics. L'analyse des paramètres de la parole a été réalisée à partir de la lecture d'un texte. Les paramètres suivants ont été déterminés : le temps maximum de phonation (TMP), le nombre de mots lus par minute (NMM) et le nombre de syllabes émises sur une seule expiration (NSE).

Analyse aérodynamique Les paramètres aérodynamiques ont été déterminés à partir du logiciel Aérophone II de Kay Elemetrics. Les paramètres pneumo-phonatoires ont été analysés en "phonation soutenue" et en "phonation confortable". En "phonation soutenue", il était demandé au patient d'émettre la voyelle /a/ à une intensité (Is) et à une fréquence fondamentale confortables le plus longtemps possible

après une inspiration maximale. Ont été déterminés : le temps maximum de phonation (TMP) et le débit d'air moyen "soutenu" (DPMs). Le quotient phonatoire (QP) a été calculé comme le rapport de la capacité vitale mesurée à partir du spiromètres type Gould II sur le temps maximum de phonation. En "phonation confortable", il était demandé au patient d'émettre la voyelle /a/ à une intensité (Ic) et à une fréquence fondamentale confortables pendant une durée de 3 à 5 secondes après une inspiration normale. Le débit d'air "confortable" a été directement mesuré (DPMc). Pour ces deux tests, l'enregistrement a été réalisé au moins 3 fois chez chaque patient. Pour la détermination des paramètres laryngés aérodynamiques, il a été demandé au patient de produire au minimum 7 fois /pi/, /pi/, /pi/...à une intensité et à une fréquence fondamentale confortable. La pression intraorale (P) a été mesurée à partir d'un petit cathéter intrabuccal simultanément au débit d'air (DPM) et à l'intensité (I). La Résistance glottique (RG) et l'Efficience phonatoire (EP) ont été calculées selon la méthode décrite par Schutte [Sch81].

Analyse morphodynamique par vidéofibroscopie laryngée avec stroboscopie. L'examen a été effectué avec un nasofibroscope Enf P3 Olympus (Pouret, Paris, France) et avec un endoscope rigide Wolf (Atmos Médical France, Marseille) sur lequel était fixé une microcaméra panasonic KS 152 (Pouret, Paris, France). Les examens étaient enregistrés sur un magnétoscope Sony U-matic 3/4 pouce, modèle VO-5800 PS (Solal, Strasbourg, France). L'ensemble des bandes vidéo a été analysé secondairement. L'analyse morphodynamique laryngée permettait de préciser en pré-opératoire et en post-opératoire, la mobilité des cordes vocales et des aryténoïdes, de visualiser la glotte "phonatoire" et "respiratoire", l'ouverture de la glotte postérieure, le mode de fermeture en phonation (fermeture glottique ou supra-glottique par adduction des bandes ventriculaires) et l'existence de vibrations muqueuses en phonation au niveau des stuctures glottiques et supra-glottiques.

Analyse statistique a été réalisée pour comparer les paramètres mesurés en préopératoire à ceux mesurés au premier mois. L'évolution entre le premier et le vingt-quatrième mois des différents paramètres a été analysée à partir d'un test de Friedman.

3. RÉSULTATS

Analyse respiratoire Après la CTP, le VIMS au premier mois était statistiquement augmenté par rapport à celui enregistré en pré-opératoire ($p = 0.01$). Il n'y avait pas de différence statistiquement significative pour les paramètres respiratoires expiratoires (Table 1). L'analyse de l'évolution par le test de Friedman a montré que la fonction respiratoire des quatre patients était stable (Table 1).

Analyse acoustique Le logiciel MDVP ne pouvait pas déterminer automatiquement les paramètres fréquentiels au premier et au troisième mois. Le test de Friedman a montré une différence statistiquement significative pour le jitter ($p = 0.04$), pour le shimmer ($p = 0.04$) et pour le HNR ($p = 0.04$) entre les $6^{ème}$, $12^{ème}$ et $24^{ème}$ mois (Table 1). Il existait une différence statistiquement significative pour les paramètres temporels entre les valeurs enregistrées en pré-opératoire et au 1^{er} mois après la CTP. Le test de Friedman a montré une différence statistiquement significative pour l'ensemble des paramètres temporels au cours de l'évolution (Table 1).

Analyse aérodynamique Les paramètres aérodynamiques sont présentés dans la Table 1 et les Figures 1 et 2. Pour l'ensemble des paramètres, il existait une différence statistiquement significative entre les valeurs enregistrées en pré-opératoire et au 1^{er} mois. Le test de Friedman a montré une différence statistiquement significative pour l'ensemble des paramètres aérodynamiques au cours de l'évolution (Table 1).

Analyse morphodynamique est représentée par les Figures 3, 4, 5 et 6.

4. DISCUSSION

En 1989, Dennis et Kashima ont décrit la cordotomie transverse postérieure au Laser CO2, et ont rapporté l'efficacité de cette intervention pour améliorer la fonction respiratoire des patients atteints de paralysie récurrentielle bilatérale [Den89]. Le concept de la CTP repose pour ces auteurs sur la préservation de la glotte phonatoire antérieure par section de la corde vocale (ligament et muscle thyro-aryténoïdien) juste en avant de l'apophyse vocale de l'aryténoïde (glotte respiratoire postérieure) (Figure 3). Laccourreye et al. ont rapporté dans une série rétrospective la rapidité et la simplicité de réalisation de cette technique, l'absence de trachéotomie, l'absence de complication postopératoire et l'amélioration subjective de la respiration sans trouble de la déglutition après CTP réalisée de façon bilatérale [Lac99]. La paralysie récurrentielle bilatérale en adduction entraîne une difficulté respiratoire sur le temps inspiratoire de la respiration. Notre étude confirme l'amélioration objective du paramètre respiratoire inspiratoire ($p = 0.01$) et la stabilité au cours de l'évolution de l'ensemble des paramètres respiratoires (Table 1). L'efficacité de la CTP sur l'amélioration de la fonction respiratoire est également démontrée par la baisse significative de la Résistance laryngée au premier mois post-opératoire ($p = 0.029$). La plupart des études ont rapporté que la qualité vocale était subjectivement bonne après CTP. L'étude de Lawson et al. [Law96] à partir de l'analyse objective du temps maximum de phonation, du quotient phonatoire, de l'intensité moyenne et de l'analyseur de fréquence à haute résolution a montré que l'ensemble de ces paramètres était proche de la normale dans une série de patients enregistrés 15.2 mois en moyenne après la

CTP. Cependant, dans notre étude, la baisse importante de la Résistance glottique en post-opératoire implique une inefficacité du sphincter laryngé et donc de la production vocale. Ce fait est également confirmé par la baisse significative de l'Efficience phonatoire au 1er mois ($p = 0.009$). L'analyse en vidéofibroscopie laryngé avec examen stroboscopique au 1er mois montre l'ouverture postérieure de la glotte avec immobilité laryngée bilatérale et l'absence de vibration au niveau des structures glottiques et supraglottiques (Figure 4). Ces données sont en accord avec l'étude de Eckel et al. [Eck94] qui ont rapporté la dégradation significative des paramètres acoustiques objectifs analysés 6 à 12 mois après cordectomie au Laser CO2. Dans notre étude, au 1er et au 3ème mois après la CTP la non détection automatique des paramètres fréquentiels (F0, jitter, shimmer) et la baisse significative de l'intensité sonore moyenne ($p = .003$) reflètent la dégradation de la fonction phonatoire. Les paramètres temporels (TMP, NMM, NSE) étaient statistiquement plus bas en post-opératoire qu'en pré-opératoire. Sur le plan aérodynamique, nos résultats montrent une augmentation significative des débits d'air phonatoire et de la pression intra orale en post-opératoire. Chez les sujets normaux contrairement à nos patients, la pression sous-glottique est positivement corrélée à l'intensité sonore. L'augmentation de la pression intra orale entre le premier mois et le troisième mois post-opératoire semble être en rapport avec le développement d'un mécanisme compensatoire lié à l'incompétence glottique. La vidéofibroscopie laryngée avec stroboscopie au 3ème mois post-opératoire met en évidence chez tous les patients un comportement supraglottique correspondant au rapprochement des bandes ventriculaires sur la ligne médiane permettant une occlusion par création d'un néo-sphincter supra-glottique (Figure 5). Ce mécanisme d'hyperfonctionnement pourrait être lié à une contraction compensatrice du muscle thyro-aryténoïdien latéral liée à l'inefficacité du muscle thyro-aryténoïdien médian (muscle de la corde vocale). Au 6ème mois post-opératoire, la vidéofibroscopie laryngée avec stroboscopie met en évidence un comportement supra-glottique avec la présence de vibrations de la muqueuse des bandes ventriculaires (Figure 5). Les perturbations des paramètres fréquentiels (F0, jitter, shimmer) et les valeurs faibles du HNR détectés à partir du 6ème mois peuvent avoir plusieurs origines : masse des bandes ventriculaires et qualités intrinsèques différentes de la muqueuse des bandes ventriculaires et création de flux aériens turbulents au niveau de la source par l'agrandissement de la glotte postérieure au Laser. Cependant, il est généralement admis que les mesures de perturbation (jitter, shimmer) n'ont de valeur que pour les voix normales ou peu dysphoniques, Woodson considère que lorsque ces mesures dépassent 10 % (ce qui était le cas chez nos patients au 6ème mois), la sensiblité et la valeur de ces paramètres doivent être considérée comme nulle [Woo98]. Le suivi des patients montre une amélioration progressive de l'ensemble des paramètres aérodynamiques. La vidéofibroscopie laryngée avec stroboscopie montre la disparition à partir du 12ème mois du comportement supra-glottique et l'apparition de vibrations de la muqueuse au niveau des cordes vocales (Figure 6). En conclusion, ce modèle montre les limites de l'analyse acoustique (F0, jitter, shimmer) et l'intérêt des mesures respiratoires, aérodynamiques et morphodynamiques pour objectiver les mécanismes physiologiques dans la voix pathologique. Les paramètres aérodynamiques et morphodynamiques détectables tout au long de l'évolution post-opératoire devraient permettre de comparer les différentes interventions par voie endoscopique (aryténoïdectomie totale ou partielle et CTP). La rééducation orthophonique devrait intégrer le phénomène d'hyperfonctionnement laryngée (cause ou conséquence de l'augmentation de pression?). Des études complémentaires sont nécessaires pour objectiver l'apport des différents types de rééducation sur le développement de ce mécanisme.

Table 1: Résultats des tests statistiques.

	Valeurs de p Pré-opératoire / 1 mois	Valeurs de p Evolution 1/24 mois
Respiratoires		
Capacité vitale	0.11	0.26
VEMS	0.05	0.43
VIMS	0.01	0.73
Peakflow	0.22	0.49
Pneumo-phonatoires		
Intensité "s"	0.002	0.003
DPMs	0.003	0.004
QP	0.003	0.001
Intensité "c"	0.004	0.002
DPMc	0.003	0.004
Aérodynamiques		
Intensité	0.003	0.003
Pression	0.001	0.005
DPM	0.003	0.003
RG	0.029	0.02
EP	0.009	0.003
Fréquentiels		
F0	ND	0.8
Jitter	ND	0.04
Shimmer	ND	0.04
HNR	ND	0.04
Temporels		
TMP	0.04	0.003
NMM	0.001	0.003
NSE	0.02	0.009

BIBLIOGRAPHIE

[Den89] Dennis D.P:, Kashima H. (1989) "Carbon dioxide laser posterior cordectomy for

treatment of bilateral laryngeal paralysis", Ann Otol Rhinol Laryngol, 98, pp. 930-934.

[Sch81] Schutte H.K. (1981), The efficiency of voice production, Groningen, Kemper.

[Lac99] Laccourreye O. et al. (1999), " CO2 laser endoscopic posterior partial transverse cordotomy for bilaterla paralysis of the vocal fold ", Laryngoscope, 109, pp. 415-418.

[Woo98] Woodson G.E.. et al. (1998), Voice analysis. Otolaryngology-head and neck surgery. C.W. Cummings, Ed, 3rd ed.Mosby-Year Book, Inc. St Louis, MO, USA, pp 1876-90.

[Law96] Lawson G. et al. (1996), "Posterior cordectomy and subtotal arytenoidectomy for the treatment of bilateral vocal fold immobility", J Voice, 10, pp. 314-319.

[Eck94] Eckel H.E. et al. (1994), "Cordectomy versus arytenoidectomy in the managemùent of bilateral vocal cord paralysis", Ann Otol Rhinol Laryngol, 103, pp. 852-857.

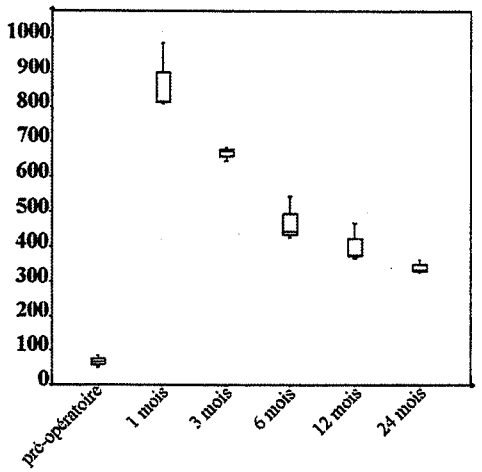


Figure 1: Evolution du débit Phonatoire Moyen (ml/s).

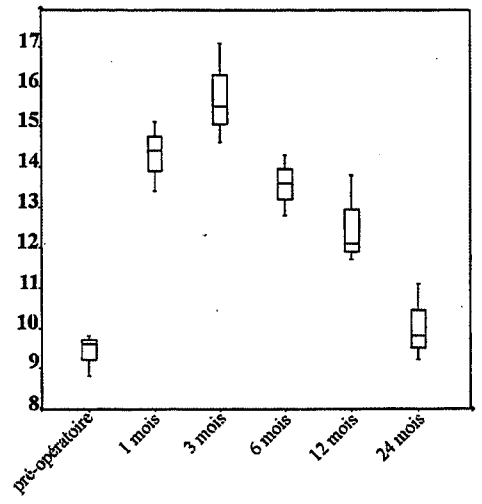


Figure 2: Evolution de la Pression Intra-orale (cmH2O)

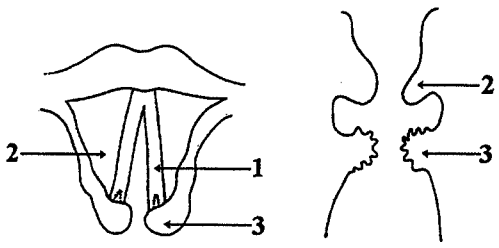


Figure 3 : Aspect en vidéofibroscopie avec stroboscopie avec coupe frontale du larynx. Immobilité des deux cordes vocales (1) et des deux aryténoïdes (3) liée à la paralysie récurrentielle bilatérale. Les deux cordes vocales sont en adduction entraînant un obstacle laryngé (augmentation de la Résistance glottique et diminution du VIMS). 2 : bandes ventriculaires.

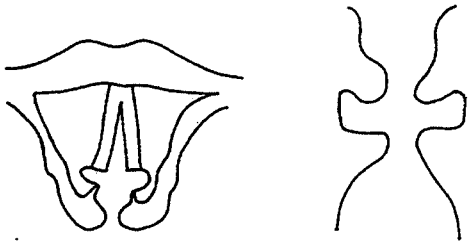


Figure 4 : Aspect du larynx à 1 mois après la CTP : visualisation du "gap" postérieur au niveau de la glotte respiratoire. Absence de vibration au niveau des cordes vocales et des structures supra-glottiques.

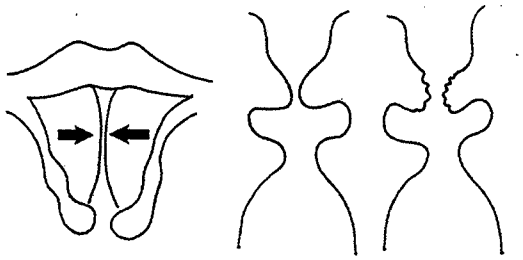


Figure 5 : Aspect du larynx au 3ème et au 6ème mois après la CTP : visualisation de la compensation par les bandes ventriculaires (comportement supra-glottique) liée à l'insuffisance du sphincter glottique. Apparition de vibrations au niveau des bandes ventriculaires au 6ème mois.

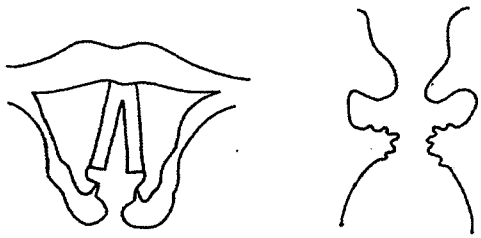


Figure 6 : Aspect du larynx au 24ème mois : disparition du comportement supra-glottique et réapparition de vibrations au niveau des cordes vocales.

Capacités phonologiques implicite et explicite chez les malvoyants

Karine Thomas, Véronique Prost, Robert Espesser, Véronique Rey

Laboratoire Parole et Langage, U.R.A. 261 CNRS & Université de Provence,
29 av. Robert Schuman 13261 Aix en Provence, cedex 1

ABSTRACT

The last ten years shown the relationship between phonological awareness and performances in reading and in writing. Our research tests the performance of normal children and visually-handicapped children aged from six to ten years old, in phonological awareness. Our hypothesis is that visually-handicapped children have a specific phonological deficit, not in their phonological system, but in their phonological awareness. The results tend to confirm this hypothesis and support the recommendation to use exercises specifically designed to develop phonological awareness.

1. INTRODUCTION

Les recherches portant sur les pré-requis pour l'apprentissage de la lecture montrent la nécessité de manipuler la langue implicitement (capacités langagières) et explicitement (capacité de segmentation explicite des unités). Nous nous sommes interrogés sur les capacités des enfants malvoyants au moment de l'apprentissage du Braille (système alphabétique tactile) afin de repérer s'ils ont une difficulté spécifique , liée à leur handicap ou non. Des travaux [Wlo.93] ont porté sur l'impact du handicap visuel dans la constitution du système phonologique des enfants malvoyants. En effet, ces travaux établissent une évaluation de la maîtrise des composants phonologiques de la parole chez les enfants malvoyants (détection de rimes, dénombrement de syllabes, association de deux mots avec un même phonème au début -les présentations des mots se faisant à l'aide de dessins-).
Nos sujets sont scolarisés et apprennent les techniques de lecture et d'écriture soit « en noir » soit en Braille. Etant donné que les capacités métalinguistiques (compétence en conscience phonologique, et répétition de logatomes) sont fortement corrélées avec les capacités en lecture, nous avons voulu connaître les capacités métalinguistiques des enfants malvoyants au moment de l'apprentissage de la lecture et de l'écriture. Notre hypothèse est double : ils ne devraient pas présenter de troubles de langage spécifiques en production spontanée, mais des difficultés dans les activités métalinguistiques car, ils sont en cours d'acquisition d'un apprentissage. Certaines de ces difficultés sont peut-être spécifiques à ces enfants, qui ne reçoivent pas (ou peu) de stimulation graphique avant l'apprentissage de la lecture et de l'écriture.

2. METHODOLOGIE

2.1 Sujets

8 enfants malvoyants (6.5 à 9.5 ans) et 10 enfants contrôles (6 à 7 ans), tous en C.P. (Cours Préparatoire) ont participé à cette étude. Les enfants malvoyants ont été recrutés dans un centre spécialisé, d'après un Q.I. normal, aucun handicap associé et une absence de trouble du langage personnel ou familial.

2.2 Expérimentation et Procédures

Les sujets ont effectué plusieurs tests :

1. Les enfants malvoyants uniquement, furent testés en parole spontanée, à partir de films réalisés dans le cadre d'une étude sur la mimo-gestuelle des émotions [Mau.98].
2. Un bilan articulatoire comprenant la répétition de syllabes simples de type CV (Consonne Voyelle), des syllabes complexes de type CCV et des mots d'usage courant en français.
3. Un bilan en conscience phonologique comprenait la recherche d'intrus : parmi trois mots, les sujets devaient désigner les deux mots pour lesquels le premier phonème consonantique, puis celui du milieu du mot, puis le dernier -exercices de rimes- étaient identiques. Cet exercice se réalisait avec des mots, puis avec des logatomes. Il comprenait aussi une épreuve de comptage phonémique (compter les phonèmes d'un mot, et compter le nombre de fois qu'un phonème donné se répétait dans un mot).
4. Une répétition de 30 logatomes.

L'ensemble des tests fut réalisé avec un support audio.

2.3 Résultats

La parole spontanée

La transcription phonétique de productions des enfants aveugles en parole spontanée a permis d'étudier la production des enfants, indépendamment de leurs compétences métalinguistiques (activité explicite sur le langage). L'analyse des productions révèlent que les sujets n'ont pas de problèmes spécifiques en conscience phonologique implicite : malgré la présence de quelques fautes phonologiques, l'analyse des énoncés ne montre pas de fautes particulières.

Le bilan articulatoire

Les résultats sont relativement homogènes. La différence entre les deux populations n'est pas significative. Il n'y aurait donc pas un déficit spécifique dans la répétition des syllabes.

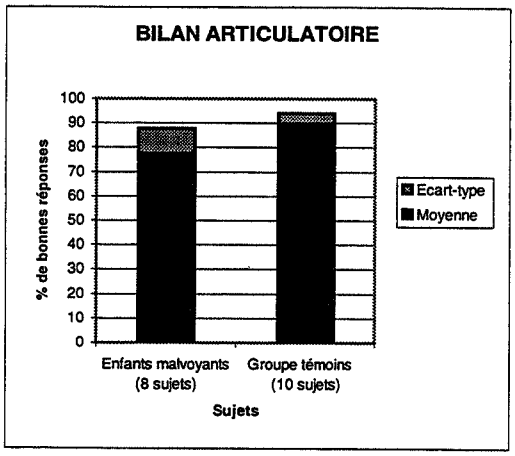


Figure 1. Résultats des tests en bilan articulatoire chez des enfants malvoyants et des témoins.

Le bilan en conscience phonologique

On peut remarquer que le pourcentage de bonnes réponses est peu élevé chez les enfants malvoyants. Ceci montre leurs difficultés à manipuler la langue de façon explicite et peut s'expliquer par leur absence de pratique en lecture et en écriture. L'analyse quantitative des fautes phonologiques commises lors de cet exercice montre qu'il y a une majorité de confusions concernant le lieu d'articulation des consonnes. Les confusions de consonnes se retrouvent dans les mots et dans les logatomes, lorsque la consonne se trouve à l'initiale ou en position intervocalique. Toutefois, la majorité des confusions a lieu dans les exercices avec des logatomes.

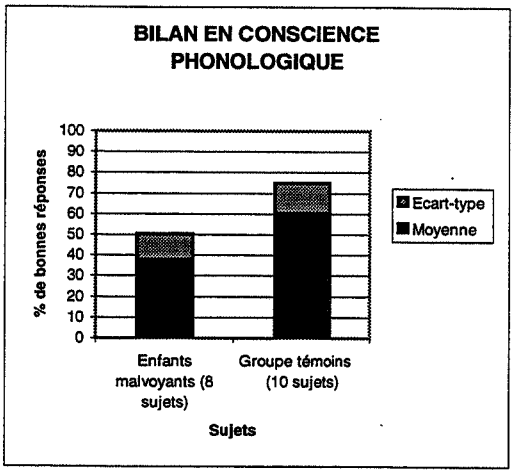


Figure 2. Résultats des tests en conscience phonologique chez des enfants malvoyants et des témoins.

Les confusions ont lieu sur les consonnes suivantes :
- dans les exercices avec des mots, confusion entre : /s/ et / /, /p/ et /t/, /b/ et /d/, /g/ et /d/, /m/ et /n/ ;
- dans les exercices avec des logatomes, confusion entre : /s/ et / /, /z/ et / /, /b/ et /d/, /p/ et /t/, /g/ et /d/, /k/ et /t/.

Nous notons également une difficulté de manipulation des voyelles en finale de mots (exercices de rimes). Il y a des confusions sur :
- l'ouverture et la fermeture des voyelles : / / et /a/, /e/ et /a/, /u/ et / / ;
- la nasalisation/dénasalisation des voyelles : /a/, /i/ et / /, et leurs correspondantes nasales : / /, / / et / /.

La répétition de logatomes

Les résultats bruts entre les deux populations sont significativement différents (Test de Wilcoxon, z=2,667, α= .008). Les enfants malvoyants présentent de grandes difficultés dans la répétition des logatomes. En outre, l'analyse quantitative des fautes phonologiques commises lors de cet exercice montre que les fautes de lieu d'articulation sont majoritaires et les sujets produisent une consonne postérieure à la place d'une consonne antérieure.

En effet, on remarque des fautes telles que :
- [fil] devient [ ] : confusion entre /f/ et / /,
- [prapu] devient [krapo] : confusion entre /p/ et /k/, ou devient [trapo] : confusion entre /p/ et /t/,
- [telysp] devient [kelyps] : confusion entre /t/ et /k/ (avec métathèse -inversion- des deux dernières consonnes), ou devient [kelyst] : confusion entre /t/ et /k/ et entre /p/ et /t/,
- [flyrp l] devient [flyrt l] : confusion entre /p/ et /t/.

La répétition de logatomes présente une spécificité phonologique chez des enfants malvoyants. Or, cette tâche relève d'une compétence plus explicite dans la manipulation des phonèmes car la voie lexicale n'est plus utilisée.

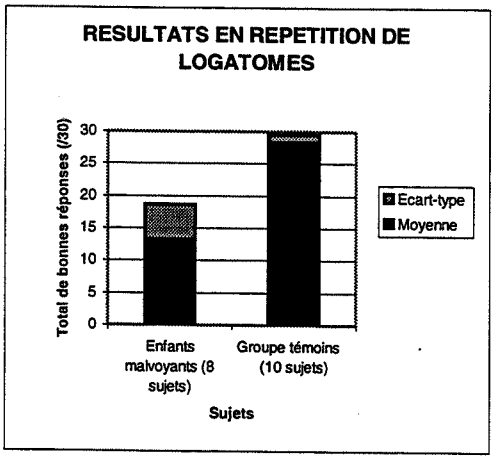


Figure 3. Résultats des tests en répétition de logatomes chez des enfants malvoyants et des témoins.

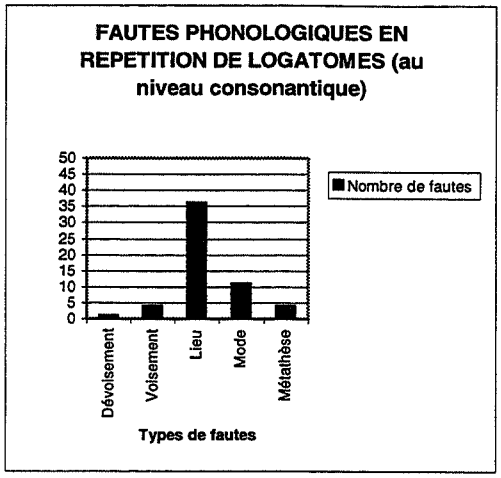


Figure 4. Fautes phonologiques commises par des enfants malvoyants lors d'une répétition de trente logatomes.

3. DISCUSSION

Ainsi comme décrit dans la littérature, les enfants malvoyants n'auraient pas une spécificité déficitaire phonologique en parole spontanée [Mil.87]. Et ceci malgré l'absence ou la quasi-absence de la lecture labiale. D'après ces travaux, l'information visuelle sur l'articulation nette donne aux enfants voyants un avantage en production. L'information visuelle dans l'articulation (mêlée à d'autres aspects comme les propriétés acoustiques et les facteurs articulatoires) joue donc un rôle dans l'acquisition de la phonologie. Toutefois, même si les enfants malvoyants sont clairement différents dans leur acquisition phonologique primaire, la différence ne vient pas nécessairement d'un désordre phonologique.
Les enfants malvoyants sont différents dans plusieurs aspects de leur développement du langage et un nombre important de ces différences provoque un retard comparé aux enfants voyants [Mil.87].

Dans les activités métalinguistiques, il s'avère que ces enfants ont une mauvaise conscience phonologique. Ceci pourrait encourager un entraînement phonologique au moment de la mise en place de l'apprentissage du Braille. Nous rejoignons ainsi les travaux de Wlomainck [Wlo.93] qui mettent en évidence qu'un entraînement du canal auditif (plus spécifiquement de la conscience phonologique) serait nécessaire car celui-ci est complémentaire du canal visuel dans l'apprentissage de la lecture.

La répétition de logatomes révèle une difficulté spécifique : les enfants malvoyants réalisent des erreurs de lieu d'articulation en faveur des consonnes postérieures. La quasi-absence de lecture labiale peut expliquer cette confusion, d'autant que les caractéristiques acoustiques entre les bilabiales et les vélaires sont très proches. Dans la perspective d'un entraînement en conscience phonologique, les exercices devraient particulièrement traiter les oppositions de lieu d'articulation.

BIBLIOGRAPHIE

[Mil.87] Mills A.E. (1987), *The acquisition of phonology in the blind child*, in Dodd B., Campbell R. (Eds), *Hearing by eye : lipreading*. Lawrence Erlbaum, Hillsdale NJ, 145-161.

[Mau.98] Maury-Rouan C. (1998), *Mimogestualité émotionnelle et coverbale chez des enfants aveugles*, in Santi S., Guaitella I., Cave C., Konopczynski G., *Oralité et Gestualité. Communication multimodal.Iinteraction*, Paris-Montréal, Paris : L'Harmattan, 81-86.

[Wlo.93] Wlomainck P. (1993), *Entraînement de la conscience phonologique en vue de l'acquisition de la lecture chez les malvoyants*, in *Revue de phonétique Appliquée*, Vol. 107, 179-190.

Croissance du conduit vocal et stratégies articulatoires en production vocalique

LUCIE MÉNARD[†], LOUIS-JEAN BOË[†] ET SHINJI MAEDA[‡]

[†]Institut de la Communication Parlée CNRS/INPG UMR 5009
38040 Grenoble Cedex 9, France
[‡]ENST/TSI CNRS, Paris France
75634 Paris, Cedex 13, France
Mél : menard@icp.inpg.fr, boe@icp.inpg.fr, maeda@tsi.enst.fr

ABSTRACT

It has been shown that the adult's vocal tract is not a uniform scaled up version of a child vocal tract. Previous work has confirmed, for children, the much shorter pharyngeal cavity compared to the front cavity. Considering these morphological differences, what are the articulatory strategies used by the speaker to produce the same vowels? Using the VLAM model [Boë97] and the concept of maximal vowel space, we determined the prototypical articulatori-acoustic realizations that a newborn (about 4 months old) would have if it were to display the same control capacities as an adult. Vowels [i], [y], [u] and [a] have been studied. Results show that articulatory adjustments are made on constriction size as well as constriction location.

1. LE PHÉNOMÈNE DE CROISSANCE

Les travaux de normalisation proposés par [Fan75] et [Nor77] montrent que les facteurs de correspondance des formants vocaliques entre l'adulte et l'enfant ne sont pas linéaires. En effet, la croissance non uniforme du conduit vocal est corroborée par les données articulatoires recueillies par [Gol80], et plus récemment par [Fit99]. Le rapport entre la hauteur du pharynx et la largeur de la cavité orale est plus important chez l'adulte que chez l'enfant, pour qui le larynx occupe une position relativement haute.

Ces différences de rapport entraînent d'importantes conséquences au plan des relations articulatori-acoustiques. Puisque les fréquences formantiques sont tributaires de la longueur des cavités, une transformation non uniforme de la longueur du conduit vocal entraînera un ajustement articulatoire afin d'atteindre un même produit acoustique F1-F2-F3. Quelles seraient les stratégies mises en œuvre par le locuteur disposant d'une morphologie différente afin de produire des voyelles typiques [i], [y], [u] et [a] ? Puisque des images cinéradiographiques de sons articulés, chez l'enfant, ne sont pas disponibles, l'exploitation d'un modèle articulatoire s'avère judicieux.

2. LE MODÈLE ANTHROPOMORPHIQUE DE CROISSANCE

Bien que les modèles articulatoires de la parole soient nombreux, force est de constater l'état lacunaire des travaux qui simulent la croissance du conduit vocal. Le modèle retenu est celui de S. Maeda (*Variable Linear Articulatory Model* [Boë97]), qui intègre les données rassemblées par l'analyse de [Gol80]. VLAM a été inséré dans un environnement (Growth) développé pour un modèle de conduit vocal d'adulte ([Boë95]), issu d'une analyse statistique d'images cinéradiographiques ([Mae79]). Ce modèle anthropomorphique est contrôlé par sept paramètres articulatoires directement interprétables en termes de degrés de liberté des articulateurs : hauteur et protrusion des lèvres, position de l'apex, du corps et du dos de la langue, hauteur de la mâchoire et position verticale du larynx. Le modèle fournit la coupe sagittale, la fonction d'aire et la réponse harmonique (formants et largeurs de bande [Bad84]) pour chaque combinaison de paramètres, ajustables selon une valeur comprise entre ± 3 écarts types. Les valeurs de fréquence fondamentale sont tirées de [Mac96].

L'évolution de la croissance est introduite dans la dimension longitudinale du conduit vocal par deux facteurs d'échelle k : l'un pour la partie antérieure du conduit vocal et l'autre pour le pharynx, la zone intermédiaire étant interpolée :

$$k\_pharynx = k(1{,}1 - 0{,}30) + 0{,}30$$
$$k\_cav\_ant = k(1{,}0 - 0{,}65) + 0{,}65$$

Ces formules ont été calibrées d'après les données rassemblées par [Gol80]. La croissance non uniforme du conduit vocal peut ainsi être simulée année par année, mois par mois. Le modèle représente donc trois types de configurations, selon le rapport des dimensions des cavités : $L_{phar} > L_{cav\_ant}$ (cas des hommes adultes), $L_{phar} < L_{cav\_ant}$ (cas des femmes) et $L_{phar} << L_{cav\_ant}$ (cas de l'enfant).

3. LA NOTION D'ESPACE VOCALIQUE MAXIMAL

3.1 Influence de la croissance

Le caractère anthropomorphique du modèle exploité permet de combiner l'ensemble des valeurs possibles des

sept paramètres de contrôle du conduit vocal et ainsi d'explorer les limites articulatori-acoustiques d'un locuteur. L'espace vocalique maximal (EVM) généré (plans F1-F2 et F2-F3) se révèle pertinent pour l'évaluation des configurations extrêmes exploitables, c'est-à-dire des voyelles [i], [a] et [u].

[Boë97] a montré l'influence de la croissance du conduit vocal sur l'EVM. En simulant l'ensemble des combinaisons des paramètres de contrôle du modèle, il ressort que l'EVM dans le plan F1-F2 est au moins aussi étendu pour le nouveau-né que pour l'adulte. Le plan F2-F3, quant à lui, est un peu plus réduit (cf. figure 1). De la naissance à la maturité, si le locuteur possédait les mêmes capacités de contrôle (sensori-moteur), les possibilités de contrastes et d'oppositions vocaliques dans le plan F1-F2 seraient les mêmes.

Par la recherche des affiliations formants-cavités, il est possible de prévoir l'influence de la dimension des cavités sur la valeur des formants. La table 1 schématise les affiliations formants-cavités retrouvées chez l'adulte (inspiré de [Mae95] et [Boë97]).

Table 1 : Affiliations formants-cavités chez l'adulte.

Voyelle	F1	F2	F3
[i]	Helmholtz arr.+constr.	arrière $\lambda/2$	avant $\lambda/2$
[a]	avant $\lambda/4$	arrière $\lambda/4$	avant $3\lambda/4$
[u]	Helmholtz av.+lèvres	Helmholtz arr. + constr.	arrière $\lambda/2$
[y]	Helmholtz av.+lèvres	arrière $\lambda/2$	avant $\lambda/2$

Rappelons que la fréquence d'un résonateur de type Helmholtz est fonction de la longueur de la cavité (L_{ca}) et de la constriction (L_{co}) mais également des aires respectives de la cavité (A_{ca}) et de la constriction (A_{co}), comme le montre la formule suivante :

$$F = (c/2\pi k)\sqrt{(A_{co}/L_{co}L_{ca}A_{ca})}$$

où c = vitesse du son (35000cm/s)

Pour les résonances de type tube uniforme fermé-ouvert, la fréquence dépend de la longueur de la cavité :

$$F = (2n-1)c/4L_{ca}$$

Les résonateurs de type uniforme fermé-fermé et ouvert-ouvert obéissent à l'équation suivante :

$$F = nc/2L_{ca}$$

où n = 1, 2, 3, 4, ...

Compte tenu des affiliations formants-cavités exposées à la table 1, l'on peut prédire qu'une transformation non uniforme de la longueur du conduit vocal, comme c'est le cas dans le passage de l'adulte au bébé, aura les conséquences suivantes :

- le contraste F2/F3 sera moins important pour [i], les deux formants devenant associés à des cavités de longueurs relativement égales;
- pour les voyelles hautes [i], [y] et [u], pour lesquelles F1 dépend des résonateurs de Helmholtz, les articulateurs recrutés agiront entre autres sur la taille de la constriction;
- les contrastes entre formants qui dépendent de résonances sous multiples de la longueur d'onde seront ajustés par des mouvements visant à antérioriser ou à postérioriser le lieu de constriction. Il est donc possible que les affiliations soient renversées.

3.2 Simulations

Un dictionnaire a été généré à l'aide de l'environnement Growth, en faisant varier chacun des paramètres de commande selon une distribution gaussienne, dans un intervalle de ±3,5 écarts types (±1 écart type pour le larynx). L'aire de constriction minimale a été fixée à 0,3 cm^2 alors que l'aire aux lèvres variait dans un intervalle allant de 0,1 cm^2 à 8 cm^2. Un ensemble de 25000 configurations articulatoires et leurs valeurs formantiques correspondantes ont été calculées, pour le nouveau-né (4 mois) et l'homme adulte de 21 ans. Les deux locuteurs types représentent respectivement des conduits vocaux de configuration $L_{phar} << L_{cav\_ant}$ et $L_{phar} > L_{cav\_ant}$. Pour chaque âge, environ 7000 points correspondaient à des productions vocaliques et ont été retenus. Nous obtenons ainsi les EVM selon différentes longueurs du conduit vocal ($L_{moyenne}$ = 17,87 cm pour 21 ans et 7,74 cm pour le bébé).

4. STRATÉGIES ARTICULATOIRES

4.1 Détermination des prototypes

Les EVM permettent d'estimer les prototypes acoustiques des voyelles suivantes, d'après les valeurs formantiques maximales et minimales :

- [i] : F3 maximal et F2 maximal
- [y] : F2 et F3 proches, et F1 minimal
- [u] : F1 et F2 minimaux (F1 et F2 focalisés)
- [a] : F1 maximal (F1 et F2 focalisés)

Dans un premier temps, les voyelles sont situées dans les limites des EVM (cf. figure 1). Il s'agit ensuite d'inverser le modèle afin de retrouver la valeur des sept paramètres articulatoires correspondants. Nous avons réalisé cette inversion grâce à une méthode utilisant la pseudo-inverse du Jacobien. Rappelons qu'il existe plusieurs solutions à l'inversion, différentes combinaisons des paramètres d'entrée correspondant au même produit (F1-F2-F3).

Ces prototypes ont été déterminés pour l'adulte par expertise [Val94]. Si l'on conserve les mêmes paramètres articulatoires de commande de l'adulte pour le bébé, on constate des décalages dans l'EVM du bébé (figure 1). La différence est particulièrement marquée pour [y] et [a], qui sont « réalisées », chez le bébé, par des timbres de type, respectivement, [ɶ], et [ʘ]. Les configurations articulatoires de l'adulte sont donc inappropriées pour un enfant qui voudrait « produire » les mêmes voyelles extrêmes. Une démarche similaire d'inversion à partir de l'EVM du bébé nous a donc permis de retrouver les paramètres articulatoires appropriés pour le bébé. La

comparaison des paramètres articulatoires inversés de l'adulte et du bébé peut être instructive à plus d'un égard.

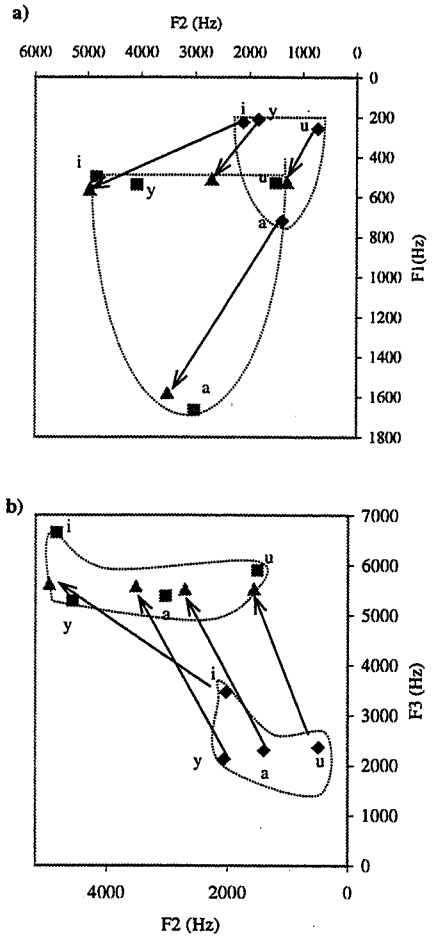


Figure 1 : Prototypes vocaliques ([i y u a]) de l'adulte (losanges), de l'enfant (carrés), et valeurs formantiques des prototypes articulatoires de l'adulte pour un conduit vocal d'enfant (triangles) (a : plan F1-F2, b : plan F2-F3).

4.2 Compensation articulatoire

Les coupes sagittales illustrées par les figures 2 à 5 représentent la configuration prototypique du bébé permettant d'obtenir la plus faible variation des paramètres articulatoires, par rapport à l'adulte. Afin de faciliter la comparaison, les paramètres du bébé et de l'adulte sont générés dans un conduit vocal de nouveau-né. Il est cependant probable de retrouver, à l'instar de [Boë *soumis à ces* JEP], des configurations différentes, impliquant une variation plus importante des paramètres articulatoires (sosies). À la lumière des affiliations formants-cavités déterminées par variation de certains articulateurs, les manœuvres articulatoires de compensation pour la différence morphologique peuvent être résumées ainsi :

- La figure 2 illustre le [i] très antérieur du bébé. Notons que cette voyelle ne nécessite que peu d'ajustements (cf figure 1), les commandes de l'homme générant un produit acoustique proche du prototype du bébé (si $A_c \geq 0{,}3\ cm^2$). La réduction importante de la cavité postérieure, par rapport à l'homme, entraîne une augmentation de F2 (affilié chez l'homme à la cavité arrière). Afin d'obtenir un contraste F2-F3 caractéristique du [i], dans un premier temps, les lèvres sont fermées, allongeant ainsi la cavité avant. Devenu affilié à celle-ci chez le bébé, F2 diminue et F3 augmente. Le dos de la langue est par la suite abaissé, antériorisant le lieu de constriction (cavité avant plus courte). Le même mouvement du dos augmente ici F3, toujours affilié à la cavité avant. Ces manœuvres des articulateurs influencent également l'aire de constriction, qui diminue F1.
- Les configurations de l'homme pour [y], selon la figure 1, génèrent une voyelle très différente chez le bébé. Les coupes sagittales de la figure 3 montrent que l'avancement du corps de la langue permet d'allonger la cavité arrière et de conserver ainsi le rapport typique de l'homme.
- La figure 4 suggère que la voyelle fermée postérieure [u], est plus antérieure chez le bébé. Afin de réaliser une focalisation F1-F2 typiquement basse de cette voyelle, un mouvement d'antériorisation du corps de la langue entraîne une diminution de la longueur de la cavité avant. Conséquemment, la longueur de la cavité arrière augmente. On retrouve à nouveau les doubles résonateurs de Helmholtz en cascade. Cette manoeuvre augmente par le fait même F3. La protrusion légèrement plus importante du bébé compense la réduction de la cavité avant, assurant un F1 bas.
- La voyelle ouverte [a], dont les formants sont affiliés à des résonateurs sous multiples de la longueur d'onde, sera réalisée avec une constriction plus antérieure, puisque les formants ne dépendent, dans ce cas, que de la longueur des cavités. Le rapport des dimensions *cavité avant/cavité arrière* sera donc rétabli par des manœuvres visant à soulever et avancer le dos et le corps de la langue. La cavité arrière devient donc plus longue, diminuant ainsi F2.

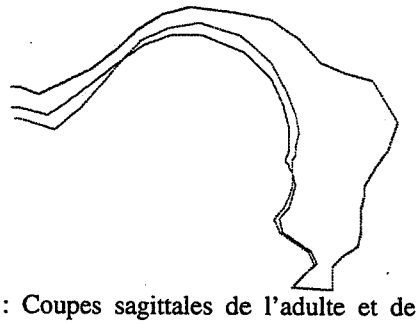


Figure 2 : Coupes sagittales de l'adulte et de l'enfant, pour [i], normalisées selon le conduit vocal du nouveau-né (trait plein : paramètres du bébé, trait tireté fin : paramètres de l'adulte).

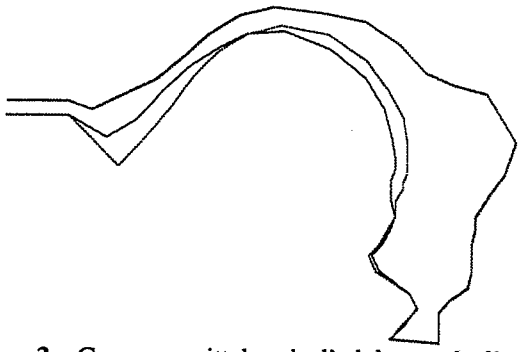


Figure 3 : Coupes sagittales de l'adulte et de l'enfant, pour [y], normalisées selon le conduit vocal du nouveau-né (trait plein : paramètres du bébé, trait tireté fin : paramètres de l'adulte).

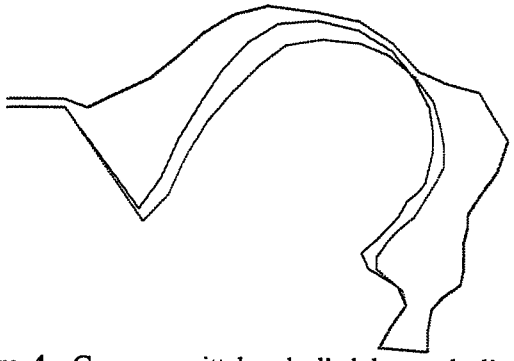


Figure 4 : Coupes sagittales de l'adulte et de l'enfant, pour [u], normalisées selon le conduit vocal du nouveau-né (trait plein : paramètres du bébé, trait tireté fin : paramètres de l'adulte).

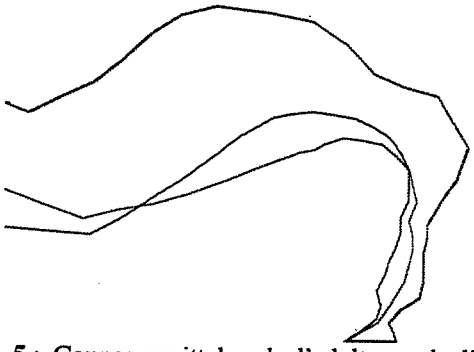


Figure 5 : Coupes sagittales de l'adulte et de l'enfant, pour [a], normalisées selon le conduit vocal du nouveau-né (trait plein : paramètres du bébé, trait tireté fin : paramètres de l'adulte).

5. Conclusion

Les résultats montrent que les stratégies permettant de compenser la différente morphologie (rapport avant/arrière) du conduit vocal du bébé donnent lieu à des commandes articulatoires différentes de l'adulte. Bien que relativement peu de changements soient nécessaires, les voyelles [i], [y], [u] et [a] sont tout de même réalisées par des manœuvres visant à modifier le lieu de constriction ou l'aire de constriction. Ces conclusions sont interprétables pour l'étude de l'ontogenèse et doivent être prises en compte pour une synthèse de parole étendue à l'enfant, de même que pour la mise au point de procédures de normalisation.

Bibliographie

[Bad84] Badin, P. et Fant, G. (1984), « Notes on vocal tract computations », *STL QPSR*, 2-3, pp. 53-108.

[Boë *soumis à ces* JEP] Boë, L.-J. (*soumis à ces* JEP), « Les sosies vocaliques. Inversion, focalisation », 23e *Journées d'Etude sur la Parole*, Aussois.

[Boë97] Boë, L.-J. (1997), « Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontogenèse et la phylogenèse », *Journées d'Études Linguistiques : « La Voyelle dans Tous ces États »*, Nantes, pp. 98-105.

[Boë95] Boë, L.-J., Gabioud, B., Perrier, P. (1995), « Speech MAPS Interactive Plant 'SMIP' », *XIIIth Int. Congr. of Phonetic Sciences*, 2, pp. 426-429.

[Boë89] Boë, L.-J., Perrier, P., Guérin, B., et J.-L. Schwartz (1989), « Maximal Vowel Space », *Eurospeech* 89, 2, pp. 281-284.

[Fan75] Fant, G. (1975), « Non-uniform vowel normalization », *STL QPSR*, 2-3, pp. 1-19.

[Fit99] Fitch, T. G. et Giedd, J. (1999), « Morphology and development of the human vocal tract : A study using magnetic resonance imaging », *JASA*, 106 (3), pp. 1511-1520.

[Mac96] Mackenzie Beck, J. (1996), « Organic Variation of the Vocal Apparatus », in W. J. Hardcastle et J. Laver, *The Handbook of Phonetic Sciences*, Oxford/Cambridge, Blackwell Publishers, pp. 256-297.

[Gol80] Goldstein, U. G. (1980), *An articulatory model for the vocal tract of the growing children*, Thesis of Doctor of Science, MIT, Cambridge, Massachusett.

[Mae79] Maeda, S. (1979) « Un modèle articulatoire de la langue avec des composantes linéaires », *10e Journées d'Etude sur la Parole*, Grenoble, pp.152-162.

[Mae95] Maeda, S. et Carré, R. (1995), *Modèles de production. École Thématique Fondements et perspectives en traitement automatique de la parole*, GDR-PRC, Communication Homme-Machine, pp. 51-72.

[Mol70] Mol, H. (1970), *Fundamental of Phonetics II : Acoustical modes generating the formants of the vowel phonemes*, Mouton, La Haye.

[Nor77] Nordström, P.-E. (1977) : « Female and infants vocal tracts simulated from male area functions », *Journal of Phonetics*, 5, pp. 81-92.

[Pet52] Peterson, G. E. et Barney, H. L. (1952), « Control method used in the study of vowels », *J. Acoust. Soc. Am.*, 24, 2, pp. 175-184.

[Roc99] Rochette, F. (1999), *Extension des possibilités de Growth [modèle articulatoire de croissance]*, Stage IUT, Départ. Informatique, Grenoble.

[Val94] Vallée, N. (1994), *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

Une base de données cinéradiographiques du français.

Alain Arnal[1], Pierre Badin[1], Gilbert Brock[2], Pierre-Yves Connan[2], Evelyne Florig[2], Noël Perez[1], Pascal Perrier[1], Pela Simon[2], Rudolph Sock[2 1], Laurent Varin[1], Béatrice Vaxelaire[2] & Jean-Pierre Zerling[2]

[1]Institut de la Communication Parlée – UMR CNRS 5009- INPG & Université Stendhal
Avenue Félix Viallet – 38031 Grenoble Cedex 01, France
[2]Institut de Phonétique de Strasbourg– Université Marc Bloch
22 Rue Descartes - 67084 Strasbourg, France.
Tél.: ++33 (0)476 57 48 25 - ++33 (0) 388 41 73 68 - Fax: ++33 (0)476 82 71 91
Mél: perrier@icp.inpg.fr – sock@ umb.u-strasbg.fr

ABSTRACT

This paper presents a preliminary version of a large X-ray database that is currently being elaborated at both the Institut de Phonétique de Strasbourg and the Institut de la Communication Parlée de Grenoble. It currently contains 4 movies that present over 2000 images. These X-ray data focus on different phonetic issues in French: juncture, nasals, and coarticulation in VCV sequences. The database contains 3 kinds of digitized data; the cineradiographic data, acoustic signals and hand-drawn sagittal contours of the vocal tract. All files are phonetically labeled and stored on CD ROMs. Management of the database is developed for Windows NT or Windows 95 with "Microsoft ACCESS", with a version for Macintosh. The data are accessed via requests in SQL language, and a user friendly interface is developed under JAVA, allowing easy formulation of requests, display of selected X-ray images and of vocal tract contours, and also listening the selected video portions.

1. INTRODUCTION : CADRE DE CE TRAVAIL

L'étude des mécanismes de contrôle de la production de la parole nécessite le recueil d'un nombre important de données tant physiologiques qu'articulatoires et acoustiques. En ce qui concerne les techniques d'acquisition de données articulatoires, il faut reconnaître qu'elles sont nombreuses et qu'elles ont fait l'objet d'efforts de développement et de diversification très intenses dans la communauté internationale au cours de la dernière décennie. Ces travaux se sont essentiellement orientés autour de la recherche simultanée d'une bonne résolution spatio-temporelle et d'une minimisation des effets négatifs potentiels sur la santé des sujets.

En effet, jusqu'au début des années 70, si on excepte les auto-observations à l'aide de miroirs d'illustres précurseurs comme Hellwag [Hel81], les techniques de bases de l'acquisition de données articulatoires reposaient sur la palatographie et la radiographie. La palatographie présentait l'intérêt d'être simple et peu onéreuse. Mais elle est restée statique jusque dans les années 70, et elle limite le champ d'observation à la zone de contact entre la langue et le palais. La radiographie impliquait le recours à une matériel médical lourd. Cependant elle offrait une représentation de la coupe sagittale du conduit vocal de la glotte jusqu'aux lèvres [Chi41], [Fan60] et, de plus, elle devint, dès la fin des années cinquante, dynamique avec l'avènement de la cinéradiographie qui permit ainsi pour la première fois l'observation attentive de mouvements des articulateurs de la parole. C'est pourquoi la cinéradiographie a connu un réel succès et a été à la base de travaux de référence sur l'articulation des sons essentiellement en français sous l'impulsion de Georges Straka [Str65] et de ses collaborateurs ([Sim67], [Roc73]), et en anglais à l'initiative de Moll [Mol60] ou Perkell [Per69].

Cependant l'exposition de sujets sains à des rayons ionisants a rapidement posé un problème déontologique, et s'est heurté à la législation de nombreux pays. Par ailleurs, la résolution temporelle de la cinéradiographie ne dépasse pas 50 images/seconde, et ceci est insuffisant pour une étude fine de l'organisation temporelle des consonnes, en particulier des consonnes plosives. Cet ensemble de raisons a incité les chercheurs à s'intéresser à de nouvelles voies d'investigation expérimentale.

Le système Xray Microbeam minimise l'exposition des sujets aux rayons ionisants en concentrant des faisceaux sur des points très localisés du conduit vocal [Wes94]. L'électropalatographie permet des mesures des contacts palataux à 100 ou 200 Hz [Har84] [Fou98]. Les techniques à ultrasons [Sto88] permettent de recueillir des données 3D à des fréquences d'échantillonnage de l'ordre de 100Hz, dans une zone limitée du conduit vocal. L'électromagnétomètre [Sch89] [Per92] mesure les déplacements de 5 ou 10 points du conduit vocal à une fréquence d'échantillonnage pouvant dépasser 1kHz. Toutes ces techniques offrent une bonne résolution spatiale, et des résolutions temporelles supérieures à la cinéradiographie. Elles ont ainsi permis d'approfondir l'étude du contrôle de la parole. Néanmoins elles souffrent d'un même défaut : aucune d'elles ne permet de recueillir en même temps des informations sur tout le conduit vocal.

L'Imagerie par Résonance Magnétique Nucléair (IRM) [Bae91] est une technique particulièrement intéressante pour pallier ce défaut. Malheureusement l'IRM dynamique n'en est qu'à ses balbutiements (moins de 10 images par seconde dans le meilleur des cas, et avec une faible résolution spatiale.

Aujourd'hui les données cinéradiographiques sont encore d'une grande utilité. Ce sont les seules à offrir aujourd'hui en même temps une résolution spatio-temporelle correcte sur l'ensemble du conduit vocal dans le plan sagittal. Elles sont à la base de l'élaboration de modèles géométriques [Mae89] et sont d'une grande utilité pour l'étude de la coordination spatio-temporelle des articulateurs de la parole [Woo97] [Vax99] [Vil99]. La législation restreint les possibilités d'acquérir de nouvelles données de ce type. Or de nombreuses banques de données existent dans différents laboratoires. Il nous semble de la plus haute importance d'en faciliter l'accès à l'ensemble de notre communauté. Munhall et ses collègues [Mun95] ont fait un magnifique travail de sauvegarde et de distribution de données cinéradiographiques réalisés en Amérique du Nord, essentiellement à l'Université Laval de Québec, sous l'impulsion de Claude Rochette.

En France, nous bénéficions d'une situation tout à fait exceptionnelle. L'Institut de Phonétique de Strasbourg a en effet accumulé, depuis la fin des années 50, plus de 50 enregistrements cinéradiographiques, et ceci sur un ensemble très large de langues. Soutenu par le *programme Ingénierie de Langues* du CNRS, nous avons donc entrepris un travail de mise en forme de ces données avec les objectifs suivants : (1) assurer la sauvegarde des données (actuellement sur films 35mm et bandes audios) par stockage sur un support vidéo de haute qualité, et par numérisation et stockage sur CDROM ; (2) apporter une valeur ajoutée par le biais de tracés sagittaux réalisés par des experts phonéticiens et montrant les limites du conduit vocal ; (3) faciliter l'accès et le traitement de ces données par leur intégration dans une base de données, et la distribution de cette base.

2. PROCÉDURE DE SAUVEGARDE ET NUMÉRISATION DES DONNEES.

2.1 Sauvegarde des films

La première étape du travail a consisté à élaborer une procédure de transfert des données, du standard cinématographique 35 mm, vers un standard plus résistant au temps, et plus facilement reproductible. Nous avons choisi le standard professionnel BetacamSP qui est une référence dans le milieu de la conservation de documents cinématographiques, et qui assure une préservation optimale de la précision des images. Nous avions deux types de films disponibles, qui différaient par la vitesse d'acquisition des images, qui était soit de 64, soit de 50 images/seconde. La vidéo restitue les images à 25 images/seconde, sous forme de trames entrelacées à 50 trames/sec. La technique que nous avons choisie, toujours dans le souci de préserver la qualité des enregistrements originaux, associe une image vidéo complète à chaque image originale (préservation de la définition spatiale) et conserve toutes les images (préservation de la définition temporelle). Cette procédure a entraîné évidemment un ralentissement de la vidéo par un facteur 2.56 (films à 64 images/seconde) ou 2 (films à 50 images/seconde). Ceci n'a évidemment aucune incidence sur les données, mais devra être pris en compte dans toute analyse temporelle ou fréquentielle ultérieure, qui reposerait sur le fichier vidéo.

À l'origine, les images et le son ont été enregistrés sur deux supports différents. L'enregistrement de tops de "synchro image" sur une des pistes de l'enregistrement audio a permis de conserver des traces de leur synchronisation originale. Ceci, associé à l'expertise des phonéticiens repérant des rendez-vous temporels entre l'image et le son (contacts labiaux pour une plosive labiale, rapprochement langue/palais pour une constrictive palatale) a permis une post-synchronisation de qualité. Le son est donc ralenti par le même facteur que la vidéo.

2.2. Réalisation de tracés sagittaux

La détermination des contours sagittaux du conduit vocal dans le plan sagittal (palais dur, palis mou, vélum, pharynx, larynx, épiglotte, langue) n'est pas chose aisée. Un certain nombre de travaux ont été développés pour élaborer des techniques d'extraction automatiques de contours [Tie94] [Lap96]. Mais la tâche reste très difficile, et nécessite le plus souvent l'assistance ou la correction d'un expert. C'est pourquoi, et ceci constitue une des particularités de notre projet, nous avons décidé d'associer aux images radiographiques brutes, un certain nombre de tracés des contours sagittaux du conduit vocal. Ces tracés offrent une plus value aux données (cf. Figure 1). Ils peuvent être exploités par un utilisateur non expert, ou servir de point de départ à de nouveaux tracés.

Toutes les images ne sont pas associées à des tracés. En effet, dans certains corpus, les séquences pertinentes sont placées au sein de phrases porteuses, dont nous n'avons pas tenu compte pour la réalisation des tracés.

2.3. Numérisation des données

Numérisation des tracés. Les tracés manuels ont été numérisés d'une part par scanner, et d'autre part via une "numérisation intelligente", qui stocke séparément les contours du palais dur, du palais mou, du pharynx, de la langue, des dents et des lèvres. Ceci permet une récupération automatique de cette information pour des traitements ultérieurs.

Numérisation des films vidéo. Les films ont été numérisés à la volée via une carte de numérisation qui génère des images numériques au standard MJPEG. Ces données ont été ensuite transformées en fichiers vidéo QuickTime, et en fichiers JPEG (images statiques). Le standard QuickTime permet une visualisation des mouvements articulatoires. Mais il correspond à un codage vidéo différentiel, et n'offre donc pas la précision spatiale requise pour un traitement ultérieur des données. C'est pourquoi nous fournissons en parallèle les images une à une codées au standard JPEG, sans perte d'information.

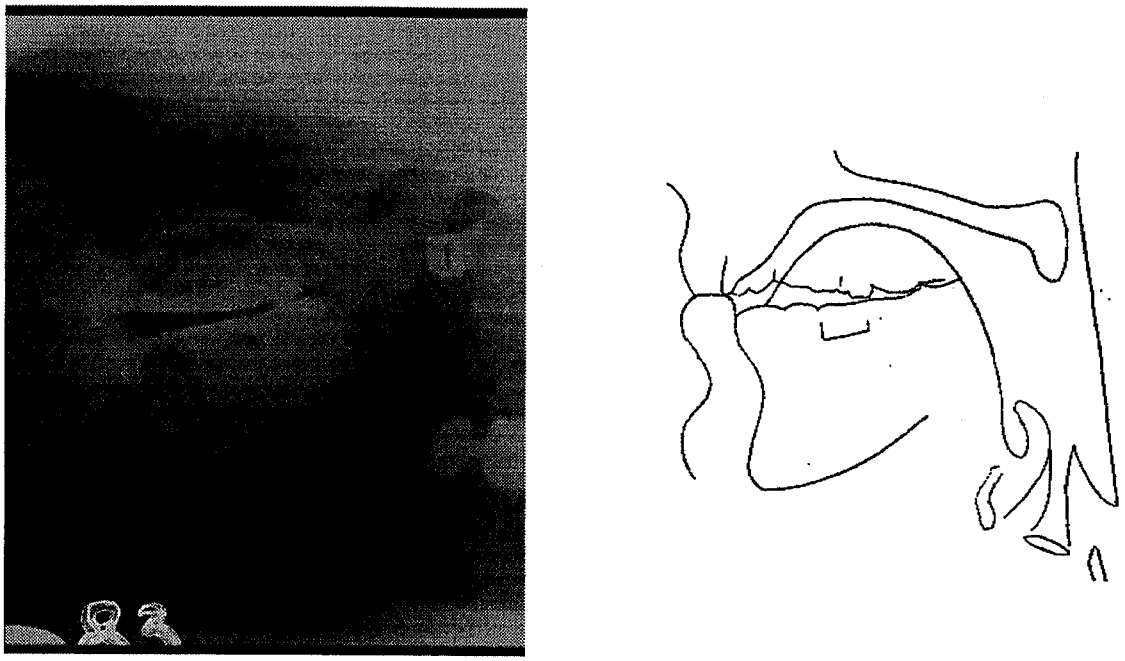


Figure 1: Radiographie de [m] dans [mi].(Figure de gauche).

Tracé des sagittaux pour la même image (Figure de gauche)

3. INTÉGRATION DANS UNE BASE DE DONNÉES.

Pour permettre une large distribution de ces données, et pour faciliter leur exploitation, nous avons posé, *a priori,* les contraintes suivantes pour la conception du système de gestion de la base de données : (1) travailler avec un SGBD relationnel d'un prix raisonnable et tournant sous des machines standards, accessibles dans tout laboratoire de parole, de type PC ou Macintosh ; (2) permettre la recherche dans la base de données à l'aide de requêtes portant sur les caractéristiques phonétiques des sons (voyelle/consonne, voisé/non voisé, ouvert/fermé...), sur l'écriture phonétique des sons, de manière isolée ou en contexte (n'excédant pas 5 phonèmes), et ceci de manière totalement transparente pour l'utilisateur.

Nous avons donc eu recours, pour les machines de type PC, au SGBD relationnel Microsoft ACCESS, et nous avons développé une interface utilisateur sous Java. Cet interface permet la formulation de requêtes à l'aide de questionnaires que l'on peut remplir de manière très ergonomique, sans que cela ne nécessite de connaissances sur les bases de données. Un manuel utilisateur a aussi été écrit en langage HTML qui permet une recherche rapide de l'information.

Pour l'administrateur de la base de données, cette interface permet aussi d'entrer de manière très simple de nouvelles données dans la base et d'en faire l'étiquetage phonétique. Cet étiquetage phonétique est grossier et détermine les zones (de l'image I à l'image I+I1), où l'on peut trouver les différents phonèmes. Les zones de transition ne sont pas étiquetées, sauf si elles sont explicitement liées à un phonème (dans le cas des liquides par exemple).

Lorsqu'un utilisateur effectuera une requête, le logiciel visualise la séquence QuickTime correspondante, et, s'ils existent, les tracés correspondants. L'utilisateur peut alors repérer le nom de la vidéo correspondante ainsi que les numéros des images retenues. Il pourra ensuite copier depuis le CDROM, les images JPEG correspondantes, ainsi que les fichiers contenant les tracés numérisés, le son de la vidéo complète, et le fichier QuickTime pour analyse ultérieure.

4. CONCLUSION.

Quatres films d'une durée de quelques minutes et associés chacun à environ 550 tracés radiographiques, ont été ainsi traités et entrés dans la base de données. Les corpus de ces films sont centrés sur les questions suivantes : *L'effet de jointure en Français* [Wio85], *Les consonnes plosives du Français* [Zer79], et *les nasales du Français* [Fla84].

La base de données sera diffusée gratuitement par l'Institut de Phonétique de Strasbourg, propriétaire des données.

BIBLIOGRAPHIE

[Bae91] Baer T., Gore J., Boyce S. & Nye P. (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. *Journal of the Acoustical Society of America, 90*, 799-828

[Chi41] Chiba T. & Kajiyama M. (1941) *The Vowel, its Nature and Structure*. Tokyo Kaseikan Pub.

[Fan60] Fant G. (1960). *Acoustic Theory of Speech Production.* Mouton, La Hague, The Netherlands.

[Fla84] Flament B. (1984) *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.

[Fou98] Fougeron C. (1998). *Variations articulatoires en début de constituants prosodiques de différents niveaux en français.* Thèse de l'Université Paris III – Sorbonne Nouvelle

[Har84] Hardcastle W. (1984) New methods of profiling lingual-palatal contact patterns with electropalatography. *Working papers Phonetics Lab.*, 4 (pp. 1-40). Université de Reading

[Hel81] Hellwag C.F. (1781). *De Formatione loquelæ.* Thèse de Médecine. Université Eberhard-Karl de Tübingen. Réédité dans *Les Cahiers de l'ICP (Bulletin de la Communication Parlée n°1).* Institut de la Communication Parlée, Université Stendhal : Grenoble, France..

[Lap95] Laprie Y. & Berger M.-O. (1996). Extraction of Tongue Contours in X-Ray Images with Minimal User Interaction. *Proceedings of ICSLP'96* (vol 1. pp.268-271).

[Mae89] Maeda S. (1989) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In Hardcastle W. & Marchal A. (Eds.) *Speech Production and Modelling* (pp. 131-149). Kluwer: Academic Publishers.

[Mol60] Moll K. (1960) Cinefluorographic techniques in speech research. *Journal of Speech and Hearing Research, 3,* 227-241.

[Mun95] Munhall, K.G., Vatikiotis-Bateson, E., & Tohkura, Y. (1995). X-ray Film database for speech research. *Journal of the Acoustical Society of America. 98,* 1222-1224.

[Per69] Perkell J.S. (1969). *Physiology of Speech Production.* Massachusetts Institute of Technology: Cambridge, Ma, USA.

[Per92] Perkell J., Cohen M., Svirsky M, Matthies M., Garabieta I., & Jackson M. (1992). Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements, *Journal of the Acoustical Society of America. 93,* 3078-3096

[Roc73] Rochette C. (1973). *Les groupes de consonnes en Français.* Presses Université Laval, Québec

[Sch89] Schönle P., Müller C. & Wenig P. (1989). Echtzeitanalyse von orofacialen Bewegungen mit Hilfe der elektromagnetischen Artikulographie. *Biomedizinische Technik, 34,* 126-130

[Sim67] Simon P. (1967) Les consonnes Françaises. Mouvements et positions articulatoires à lumière de la radiocinématographie. Paris: Klincksieck

[Sto88] Stone M., Shawker T., Talbot T. & Rich A. (1988). Cross-sectional tongue shape during vowels. *Journal of the Acoustical Society of America, 83,* 1586-1596.

[Str65] Straka G. (1965). *Album Phonétique.* Presses de l'Université Laval, Québec

[Tie94] Tiede M. & Vatikiotis-Bateson E. (1994). Extracting articulator movement parameters from a videodisc-based cineradiographic database. Proceedings of ICSLP 94 (pp.45-48)

[Wes94] Westbury J.R., Turner G. & Dembovski J. (1994) *X-ray microbeam speech production database users' handbook.* Waisman Center, Université du Wisconsin.

[Vax99] Vaxelaire B., Sock R., Bonnot J.F. & Keller D. (1999). Anticipatory labial activity in the production of French rounded vowels. *Proceedings of ICPhS 99* (Vol. 1., pp. 53-56).

[Vil99] Vilain A., Abry C.& Badin P. (1999). Motor equivalence evidenced by articulatory modelling. *Proceedings of Eurospeech99* (Vol.1, pp. 169-172)

[Wio85] Wioland F. (1985) *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.

[Woo79] Wood S.A.J. (1979). A radiographic examination of constriction location for vowels. *Journal of Phonetics, 7,* 25-43

[Woo79] Wood S.A.J (1997). A cinefluorographic study of the temporal organization of articulator gestures: Examples from Greenlandic. *Speech Communication, 22,* 207-225.

[Zer79] Zerling J.-P. (1979) *Articulation et coarticulation dans des groupes occlusive-voyelle en français. Etude cinéradiographique et acoustique : contribution à la modélisation du conduit vocal.* Doctorat 3° Cycle, Institut de Phonétique, Université de Nancy II.

Etude expérimentale de l'écoulement d'air dans des cordes vocales asymétriques. Application aux voix pathologiques.

Coriandre Vilain, Xavier Pelorson, Yann Falevoz [(1)], Mico Hirschberg[(2)]

(1) Institut de la Communication Parlée

INPG, 46 av. Félix Viallet -38031 Grenoble cedex 01

Email: cvilain@icp.inpg.fr

tel : 04-76-57-47-13, fax :04-76-57-47-10

(2) Technical University of Eindhoven

W&S, Postbus 512 5600 MB Eindhoven The Netherlands.

ABSTRACT

In this study, we present our experimental research on the interactions between the flow and the vocal folds in the case of asymmetric geometries of the folds. Such a case can appear in pathological voices where one fold is paralysed or removed by surgery. First we present the apparatus used to model the vocal tract. Then the experimental results are shown. In a first step, we discuss the possibility of flow asymmetry even in a symmetric channel (Coanda effect). In a second step, we consider asymmetric geometries. A comparison between theory and experiment is presented and tends to show that the current assumption of a symmetric flow model is not accurate in the case of asymmetric vocal folds.

1. INTRODUCTION

D'un point de vue physique, les sons de parole proviennent principalement d'une interaction entre l'air expiré par les poumons et le conduit vocal. Dans le cas des sons voisés, ce sont les cordes vocales qui - sous l'action de l'air - oscillent et produisent un son. Les premières tentatives de modélisation des cordes vocales ont d'abord cherché à en simplifier les caractéristiques géométriques et physiologiques. Il en a résulté le fameux modèle à deux masses ([Ish72]) qui considère chaque corde comme un système de deux masses couplées par un ressort et un amortisseur. A ce modèle mécanique simple est appliquée une force exercée par l'écoulement d'air qu'il faut donc aussi modéliser. Les travaux de Pelorson et coll. ([Pel95]) ont montré que le modèle géométrique d'Ishizaka et Flanagan (deux masses de forme rectangulaire) induisait un comportement de l'écoulement irréaliste en parole (phénomène de vena contracta, point de séparation fixe) et que ce modèle devait être reconsidéré. Dans la continuité de ces travaux, l'étude que nous présentons dans cet article a pour but d'évaluer l'importance de l'interaction entre l'écoulement d'air et les cordes vocales dans le cas plus complexe d'une géométrie asymétrique. Dans la pratique, une telle asymétrie peut être causée par une paralysie ou par l'ablation de l'une des cordes. Dans ce cas, il apparaît important de disposer de modèles théoriques simples qui permettent d'évaluer l'effet d'une telle paralysie sur la production de parole. De nombreux chercheurs se sont penchés sur le problème ([Ish76], [Ste94]) et dans la plupart des études, ils se sont concentrés sur l'asymétrie du modèle mécanique (géométrie et caractéristiques internes des cordes) en supposant le modèle d'écoulement identique à celui correspondant au cas des géométries symétriques. Nous allons dans cet article tenter d'évaluer la validité de cette hypothèse. Pour cela, nous avons réalisé une étude expérimentale sur maquette destinée à mesurer les caractéristiques de l'écoulement (pression et vitesse) au niveau de modèles de cordes vocales asymétriques. Dans un premier temps, nous allons décrire le dispositif expérimental utilisé, puis nous présenterons et analyserons les résultats obtenus.

2.DISPOSITIF EXPÉRIMENTAL

2.1 Caractéristiques générales

La maquette développée à l'ICP comporte en son extrémité les répliques de cordes vocales au niveau desquelles sont effectuées les mesures de pression (capteurs Kulite XCS-093) et de vitesse (anémométrie par fil chaud) (figure 1). Les pressions sont mesurées en trois endroits: en amont de la glotte (pression subglottique Psub), et de part et d'autre de la glotte (pressions glottiques P1 et P2). Les vitesses peuvent être mesurées par un fil chaud dans tout le canal (une règle graduée au demi-millimètre mesure la position du fil chaud suivant y, une règle graduée au millimètre mesure la position en x).

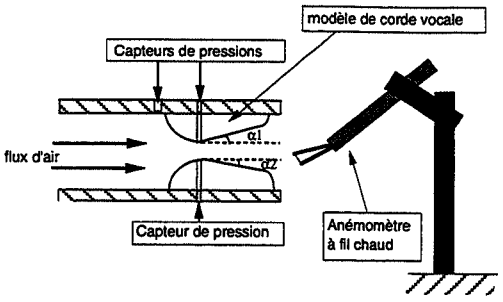


Figure 1 : schéma de la zone de mesure.

Il est important de noter que ces modèles de cordes vocales sont rigides (ceci afin de contrôler plus facilement les paramètres physiques de l'expérience). Ils représentent une configuration géométrique divergente. Cette configuration a lieu lors de la phase de fermeture des cordes. Elle est la plus intéressante à étudier car elle présente la complexité maximale en ce qui concerne l'écoulement. D'autre part, nous pouvons nous affranchir du caractère rigide du modèle de cordes en lui faisant parvenir un écoulement oscillant (alors que dans le cas réel, les cordes vocales mobiles sont soumises à une pression subglottique constante en première approximation). Cet écoulement oscillant est réalisé au moyen d'un dispositif spécial appelé tuyau collabable (figure 2).

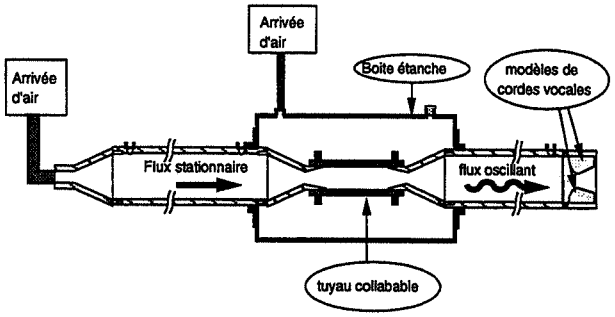


Figure 2 : Vue globale du dispositif expérimental. Rôle du tuyau collabable

Ce dispositif, utilisé initialement pour l'étude de l'écoulement du sang dans les veines de mammifères (voir par exemple les travaux de Pedley [Ped80]) consiste en un tuyau souple inséré dans une boite étanche dont on peut contrôler la pression. L'effet de la pression dans la boite - qui tend à fermer le tuyau - est contrebalancé par l'effet de l'écoulement dans le tuyau. Pour certaines valeurs de pression, le tuyau auto-oscille et génère donc un débit pulsé. L'intérêt d'un tel dispositif réside dans la possibilité de reproduire des débits proches de ceux obtenus lors de la phonation. En fait, le tuyau collabable a lui même un comportement similaire à celui des cordes vocales. La tension musculaire de celles-ci est remplacée par la pression d'air dans la boite étanche (qui tend à fermer le tuyau) et l'air passant par le tuyau joue le même rôle que l'air expiré par les poumons. La figure 3 présente un exemple de mesure de vitesse au niveau des modèles de cordes vocales dans le cas d'un écoulement oscillant produit par le tuyau collabable.

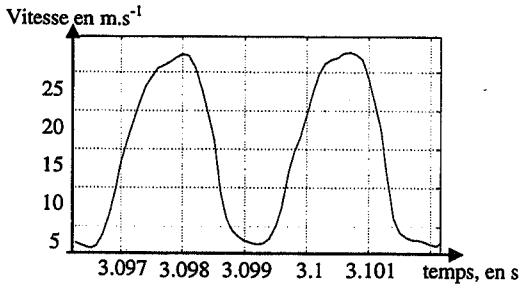


Figure 3 : Exemple de débit oscillant généré par le tuyau collabable.

2.2 Configurations géométriques et caractéristiques de l'écoulement

Chaque modèle de corde vocale est caractérisé par son angle de divergence α compris entre 0° et 40° (par pas de 10°). Une configuration géométrique donnée est alors définie par un couple d'angles (α_1,α_2) ainsi que par la hauteur minimale de la constriction (glotte): h_g. Dans un premier temps nous avons étudié des configurations symétriques puis ensuite asymétriques. Différents angles globaux de divergence $(\alpha_1+\alpha_2)$, ainsi que différents degrés d'asymétrie $(|\alpha_1-\alpha_2|)$, ont pu être comparés.

Concernant l'écoulement, nous avons réalisé plusieurs types de mesures: d'une part en régime stationnaire (on laisse le régime permanent s'établir après une ouverture de vanne), d'autre part en régime instationnaire (ouverture de vanne ou tuyau collabable). Chaque mesure sera donc caractérisée par sa géométrie (symétrique, asymétrique, hauteur à la constriction) et par le type d'écoulement considéré.

3. EFFETS DE L'ASYMÉTRIE DES CORDES SUR L'ÉCOULEMENT.

3.1 Résultats expérimentaux

Cas des géométries symétriques: Nous avons dans un premier temps vérifié l'influence d'une géométrie symétrique sur le jet. Il apparaît que dans le cas d'écoulements stationnaires, contrairement à une idée intuitive, le jet n'est généralement pas symétrique. L'effet Coanda, responsable de l'asymétrie, résulte d'une instabilité initiale du jet qui tend à se coller sur l'une des parois. Cet effet est lié au phénomène de séparation de l'écoulement, il apparaît par conséquent uniquement dans des configurations divergentes. La figure 4 présente une mesure de profil vertical de vitesse dans le cas d'une géométrie symétrique d'angle (20°,20°) et d'un écoulement stationnaire. Le fil chaud est positionné à l'extrémité des cordes. La hauteur de constriction h_g est égale à 1.25 mm.

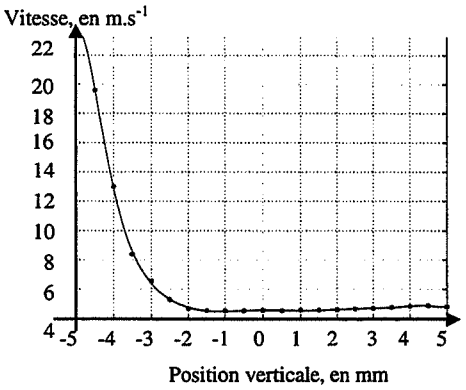


Figure 4: Profil de vitesse dans un écoulement stationnaire pour une configuration géométrique symétrique (20°,20°)

L'asymétrie de l'écoulement due à l'effet Coanda est donc clairement observable sur ce profil.

Une des caractéristiques de cet effet est son délai avant apparition. On peut voir, dans le cas d'une ouverture de vanne, qu'au tout début les pressions glottiques sont égales et qu'elles diffèrent ensuite. Si l'on considère le signal de vitesse, on aperçoit un brusque changement au moment où cet effet apparaît (figure 5).

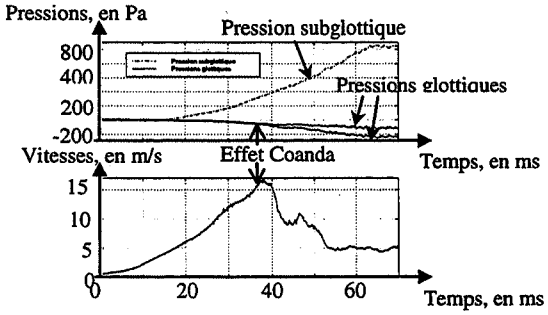
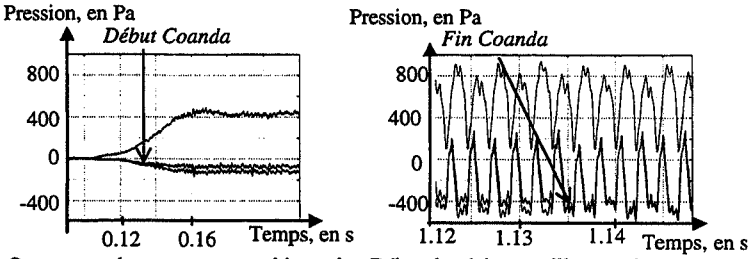


Figure 5: Apparition de l'effet Coanda (flèche). Visualisation sur les signaux de pression et de vitesse.

La question qui se pose dans le cas d'une modélisation de l'écoulement au niveau des cordes vocales est de savoir si un tel effet a lieu lors de la phonation. Si tel est le cas, il peut être nécessaire d'en tenir compte dans le modèle théorique. Cependant, pour se rapprocher des conditions de la phonation, il est important de prendre en compte l'aspect non stationnaire de l'écoulement. Il apparaît alors généralement que dans ce cas, l'effet Coanda n'a pas lieu (figure 6). La raison en est que le temps nécessaire à son apparition (typiquement 20 ms) est plus long que la période d'oscillation des cordes (10 ms) voire même beaucoup plus long que la phase de fermeture des cordes pendant laquelle cet effet peut se produire. L'effet Coanda n'a donc pas le temps d'apparaître. (voir par exemple [Vil99])

NB : Dans nos expériences, l'effet Coanda ne disparaît pas systématiquement dans le cas d'écoulements oscillants. La raison que nous avançons est que pour certains régimes d'oscillation, le tuyau collabable ne se ferme pas complètement (ce que nous pouvons voir sur



Ouverture de vanne, apparition de l'effet Coanda après un délai de 20 ms.

Début du régime oscillant, après quelques périodes d'oscillation, disparition de l'effet Coanda.

Figure 6: Apparition et disparition de l'effet Coanda dans un même signal.

les mesures de débit).L'effet Coanda - qui apparaît lors de l'ouverture initiale de la vanne - n'est donc pas réinitialisé à chaque fermeture du tuyau lors de l'oscillation. Dans le cas de la phonation, une fermeture plus complète des cordes a lieu.

Cas des géométries asymétriques : Voici un exemple de mesures de pressions réalisées sur des cordes asymétriques pour des régimes d'oscillation similaires.

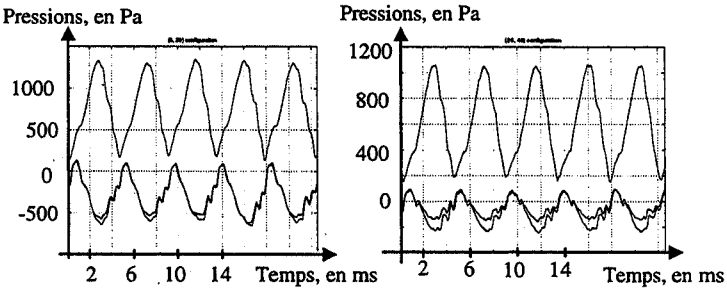


Figure 7: Mesures de pression réalisées sur deux configurations. A gauche: (α_1,α_2) =(0°,20°), à droite: (α_1,α_2)= (20°,40°).

Nous avons évalué l'asymétrie de l'écoulement en définissant un paramètre d'asymétrie $A = 100 * \left| \frac{P_2^{RMS} - P_1^{RMS}}{\min(P_2^{RMS}, P_1^{RMS})} \right|$. Dans l'exemple de la figure 7, nous avons dans le cas de la configuration (0°,20°): A= 7% (faible asymétrie), alors que dans le cas de la configuration (20°,40°): A = 58% (forte asymétrie). Il semble donc à première vue que l'asymétrie A soit directement dépendante de l'angle d'ouverture global $(\alpha_1+\alpha_2)$. Une étude systématique ([Vil99]) nous a cependant montré que cette conclusion devait être modérée. Des configurations de type (α_1,α_2) = (0°, X°) où X = (10°, 20°, 30°, 40°) entraînent par exemple une faible asymétrie du jet alors que d'autres moins divergentes peuvent causer une plus forte asymétrie (ex : (0°, 40°) comparé à (10°, 20°)).

3.2 Comparaison entre théorie et expérience

Pour tenter de valider l'hypothèse de symétrie du modèle d'écoulement dans le cas de géométries asymétriques, nous avons comparé les mesures expérimentales (configurations asymétriques) avec les modèles théoriques dont nous disposons à l'heure actuelle et qui prennent en compte des géométries

symétrique. Ces modèles se basent sur une théorie de couche limite exprimée sous une forme intégrale (équation de Thwaites, [Sch99]). Ils permettent généralement un très bon accord théorie/expérience dans le cas d'écoulements stationnaires sur des géométries symétriques divergentes (mais ne reproduisent pas l'effet Coanda). L'instationnarité de l'écoulement est quant à elle plus difficile à prendre en compte. Cependant l'utilisation d'une théorie quasi-stationnaire dans le cas d'écoulement faiblement instationnaires donne déjà des résultats satisfaisants.

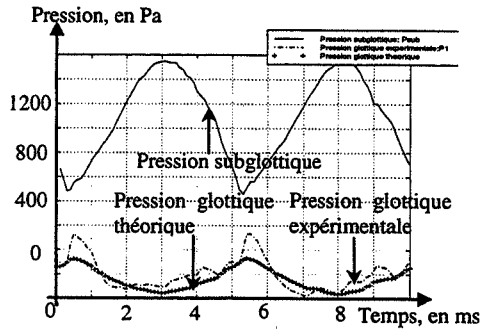


Figure 8: Comparaison expérience (instationnaire)/ théorie (couche limite quasi-stationnaire) pour une géométrie symétrique (10°,10°)

Sur la figure 8, nous avons comparé la pression glottique calculée par le modèle théorique avec la pression glottique mesurée. La pression subglottique, l'angle de divergence et la hauteur h_g servent de paramètre d'entrée pour le calcul théorique. Hormis une prise en compte encore inexacte de l'instationnarité, nous pouvons voir que la prédiction théorique suit en moyenne la mesure expérimentale.

Voici maintenant un exemple de comparaison entre une mesure réalisée sur une configuration asymétrique (20°,40°) et la prédiction théorique faite sur un modèle (30°,30°).

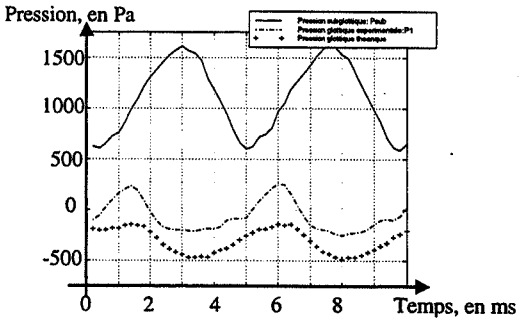


Figure 9: Comparaison expérience (écoulement instationnaire, géométrie asymétrique d'angle (20°,40°)/ théorie de couche limite quasi-stationnaire prenant en compte une géométrie symétrique (30°,30°).

Nous voyons maintenant que la différence entre théorie et expérience est plus systématique que dans le cas de la figure 8. Il semble donc que l'hypothèse de fluide symétrique dans le cas de géométries asymétriques ne soit pas validée par notre modèle théorique.

CONCLUSION

Nous avons effectué un ensemble de mesures expérimentales sur maquette dans le but de mieux comprendre les phénomènes d'interaction entre géométrie des cordes vocales et écoulement. Notre attention s'est portée sur les configurations géométriques divergentes soumises à des écoulements stationnaires ou instationnaires. Dans le cas de géométries symétriques, nous avons mis en évidence l'existence d'une instabilité du jet qui tend à le rendre asymétrique : l'effet Coanda. Cet effet disparaît généralement dans le cas d'écoulements oscillants. Nous avons ensuite concentré notre étude sur les configurations géométriques asymétriques pour lesquelles les modèles d'écoulement sont peu nombreux. Une comparaison des résultats expérimentaux avec une théorie de couche limite quasi-stationnaire semble aller contre l'hypothèse (faite par de nombreux auteurs) d'un écoulement symétrique dans une configuration asymétrique.

BIBLIOGRAPHIE

[Ish72] : Ishizaka K., Flanagan J.L.(1972) " Synthesis of voiced sounds from a two-mass model of the vocal cords. " Bell Syst. Tech. J., vol 51 (6), 1233-1268.

[Pel95] : Pelorson X., Hirschberg A., Wijnands A.P.J., Baillet H. (1995) " Description of the flow through in-vitro models of the glottis during phonation. " Acta Acustica ,vol. 3, 191-202

[Ish76] : Ishizaka K., Isshiki N.(1976) " Computer simulation of pathological vocal-cord vibration ." J.Acoust.Soc.Am., vol. 60, n° 5, 1193-1198

[Ste94] : Steinecke I., Herzel H. (1994) "Bifurcations in a asymmetric vocal fold model", NCVS Status and Progress Report, vol 6, 19-31.

[Ped80] : Pedley T.J. (1980) "The fluid mechanics of large blood vessels", Cambridge University Press.

[Vil99]: Vilain C., Pelorson X., Thomas D. (1999) " Effects of an induced asymmetry on the flow through the glottis in relation to voice pathology. " proc. of the International Workshop on models and analysis of vocal emissions for biomedical applications. Florence, 1-3 septembre 99

[Sch99]: Schlichting H, Gersten K. (1999) "Boundary Layer Theory" (8ème édition), Springer Verlag.

Remerciements : Ce travail à été financé en partie par le CNRS dans le cadre de la coopération internationale (projet EUR99/NDL) ainsi que par le programme d'action intégrées « Van Gogh » du Ministère des Affaires Etrangères.

Production de parole après traitements de cancers de la cavité endobuccale.

† Christophe Savariaux, Pascal Perrier,
‡ Jacques Lebeau, Gustavo Magaña & Chloé Dorange-Pattoret

† Institut de la Communication Parlée, UMR CNRS 5009 – INPG & Université Stendhal
46 avenue Félix Viallet – 38031 Grenoble Cedex 01 – France
‡ Service de Chirurgie Plastique et Maxillo-Faciale, CHU Grenoble, BP 217
38043 Grenoble Cedex 09 – France

ABSTRACT

The ability to speak after tongue surgery including partial glossectomy or mandibulectomy and tongue reconstruction is studied. Fourteen French speaking patients pronounced a set of French vowels and VCV sequences. Data analysis was based on formant frequencies for the vowels, and on the burst spectrum, the VOT and the hold duration for the consonants. Different degrees of impairment were observed for the vowels, and they were used to classify the patients into 3 groups. Consonant analysis was then carried out separately for each group. Results are interpreted in terms of speech representations and vocal tract perturbations and speech quality after surgery.

1. INTRODUCTION

Cet article présente les résultats d'une évaluation de la production de parole chez des patients ayant subi une opération de la cavité endobuccale avec perte de substance puis reconstruction. Notre démarche diffère de celles qui sont classiquement publiées dans la littérature [Del97], en ce sens que nos analyses sont centrées non pas sur la capacité des patients à communiquer par la parole, mais sur leur habileté à produire les configurations géométriques du conduit vocal nécessaires à l'articulation de la parole. Pour cela nous nous sommes appuyés sur une évaluation quantitative fondée sur la mesure de paramètres pertinents extraits du signal acoustique de parole. Deux objectifs ont guidé cette démarche, celui de la compréhension des mécanismes de compensation mis en jeu par les locuteurs, et celui de l'étude de l'impact sur la parole des différentes techniques chirurgicales et de reconstruction appliquées aux patients.

2. PROTOCOLE EXPÉRIMENTAL

2.1 Corpus

Le corpus utilisé pour cette expérience a été élaboré pour évaluer les capacités des patients à :

- produire des voyelles tenues sans mouvement ; les 10 voyelles du français ont été étudiées.
- produire des transitions sans contrainte temporelle forte ; des séquences V-V ont été étudiées.
- contrôler précisément leur articulation tout en assurant une coordination temporelle entre la langue et les cordes vocales ; une série de séquences VCV a été étudiée, où le degré de complexité de la tâche augmentait depuis des séquences du type /aCa/ vers des séquences du type /aCi/ et /iCa/ qui impliquent un large mouvement de la langue.

2.2 Procédure

Quatorze patients de langue maternelle française ont été enregistrés au CHU de Grenoble. Ces patients ont subi une glossectomie partielle ou totale avec reconstruction, ou une mandibulectomie partielle. Les enregistrements ont été faits pour chaque patient lors d'une unique session, et dans un délai allant de 3 semaines à 3 ans après le jour de l'intervention chirurgicale. Le corpus a été présenté, sans consigne temporelle, dans l'ordre suivant : les voyelles isolées, les transitions V-V puis les séquences VCV.

2.3 Analyse des données

Le signal acoustique a été numérisé à 20 kHz. Pour les voyelles, l'analyse a été fondée sur les fréquences des 3 premiers formants déterminées par une analyse LPC. Conformément à une démarche classique, la variation de F1 a été interprétée comme une mesure du degré d'aperture du conduit vocal, alors que la variation de F2 a permis d'estimer la position relative avant/arrière de la langue. L'ordre de grandeur de F3 a été exploité pour évaluer la capacité du sujet à produire des articulations extrêmes dans la région palatale : un F3 élevé correspond à une articulation très antérieure (type [i]) et un F3 bas est associé à une articulation très postérieure (type [u]). Les consonnes ont été analysées dans le domaine temporel par la mesure du VOT et de la durée de la tenue consonantique, et dans le domaine spectral via le spectre de l'explosion (plage de fréquence dominante, forme de l'enveloppe spectrale) et les transitions formantiques. Ces données ont été comparées à des mesures effectuées dans les mêmes conditions sur 2 locuteurs normaux, ainsi qu'à des données publiées dans la littérature. Ici, l'étude est limitée aux plosives non-voisées [t, k].

3. ANALYSE DES VOYELLES

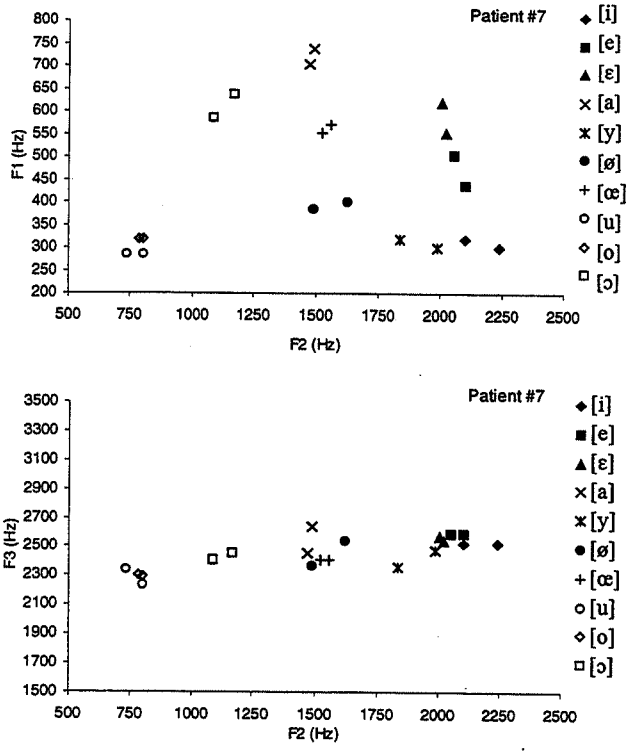
3.1 Observations

L'analyse des productions vocaliques des 14 patients a permis de les classer en 3 groupes.

Le groupe 1, représenté par le locuteur #2, comprend 5 patients qui ont conservé la distribution classique des voyelles au sein du triangle vocalique, que ce soit en condition isolée ou en contexte. Il semble donc que, pour

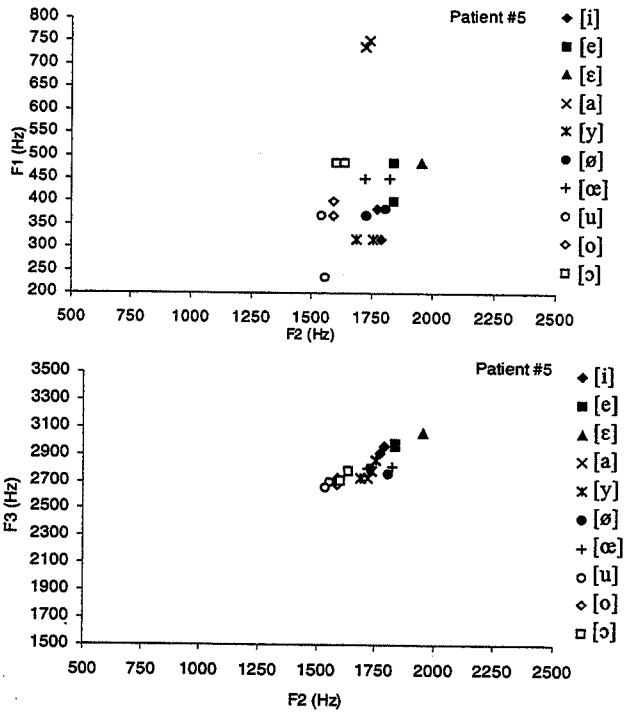
ces locuteurs, la chirurgie n'ait pas généré de gêne significative de la mobilité des articulateurs.

Le groupe 2 correspond à 6 patients qui semblent éprouver des difficultés à coordonner des mouvements simultanés avant/arrière et haut/bas de la langue et/ou à discriminer les voyelles fermées. C'est le cas du patient #7 dont la figure 1A montre les réalisations vocaliques dans le plan F1/F2 : on n'observe pas de difficultés notoires à produire les voyelles extrêmes [i, a et u], mais la position des autres voyelles à l'intérieur du triangle vocalique n'est pas conforme aux observations classiques. On note ainsi une position inhabituelle des voyelles centrales ([ɛ], [o]) dont l'articulation nécessite une coordination des gestes haut/bas et avant/arrière de la langue. La figure 1B confirme ces difficultés dans le plan F2/F3 : la plage de variation de F3 est très faible, se limitant, toutes voyelles confondues, à l'intervalle 2,2 – 2,6 kHz, suggérant que les positions extrêmes de la langue ne sont pas atteintes. Les séquences V-V sont correctes.



Figures 1A et 1B : Représentation des voyelles du patient #7 dans les plans F1/F2 et F2/F3.

Les 3 autres locuteurs, formant le groupe 3, montrent d'importantes difficultés pour produire des distinctions entre les 10 voyelles, comme en attestent les figures 2A et 2B pour le patient #5. F2 est cantonné entre 1,5 et 2 kHz, ce qui suggère une faible amplitude du geste avant/arrière de la langue. De plus, les voyelles sont mal séparées dans la dimension du premier formant, si on excepte la voyelle extrême [a]. Nous interprétons cela comme la conséquence de difficultés à contrôler différents degrés d'aperture. Toutefois il est intéressant de noter que l'ordre dans le degré d'aperture a été préservé. La faible dispersion des mesures dans le plan F2/F3 confirme que le patient ne parvient pas à déplacer sa langue dans des positions extrêmes (figure 2B).



Figures 2A et 2B : Représentation des voyelles du patient #5 dans les plans F1/F2 et F2/F3.

3.2 La nature de la chirurgie et ses effets

Cette classification est-elle cohérente avec la nature des interventions chirurgicales ? Pour répondre nous allons détailler les interventions subies par les sujets.

Le patient #2 (groupe 1) a subi 2 mandibulectomies à 5 ans d'intervalle, puis une radiothérapie. La reconstruction de l'os mandibulaire n'a pas été complète et sa partie gauche n'est pas solide. La langue est restée anatomiquement inchangée.

Le patient #7 (groupe 2) a subi l'ablation d'une large partie du plancher de la bouche ainsi que d'une petite partie de la racine de la langue. Le plancher de la bouche a été reconstruit avec un lambeau libre du muscle péroné. Ceci est associé à une sclérose des muscles de la langue.

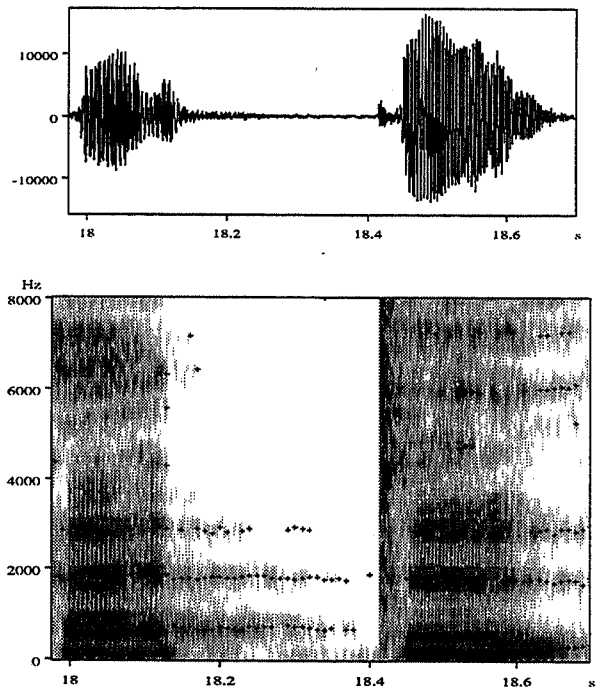
Le patient #5 (groupe 3) a suivi une Curiethérapie qui a induit une fibrose des muscles de la langue. Trois ans plus tard, il a subi une hémi-glossectomie gauche, depuis la racine jusqu'au bout de la langue. Une reconstruction a été faite à partir d'un lambeau du muscle grand dorsal.

Ainsi il apparaît que l'impact de la chirurgie sur la langue va en croissant depuis le patient du groupe 1 à celui du groupe 3. La capacité qu'ont les patients à produire les voyelles décroît dans le même ordre. Ceci confirme le bien fondé de notre stratégie qui consiste à faire une première évaluation du degré de handicap sur la base des productions vocaliques.

4. COORDINATION DES MOUVEMENTS ARTICULATOIRES

Rappelons que seule l'analyse des plosives non-voisées [t, k] est faite dans ce papier.

L'étude du patient #5 témoigne de grosses difficultés à produire les consonnes. Pour le [t] de [ita] ou [ati], le spectre du burst est quasi identique à celui des locuteurs normaux, couvrant une large plage de fréquence avec un maximum entre 5 et 7 kHz. Après l'explosion apparaît un bruit de friction long et intense, surtout en contexte [ati] (figures 3A et 3B)



Figures 3A et 3B : Signal et sonagramme produit par le patient #5 durant la séquence [ati].

Pour [ata] le spectre du burst est différent de celui des productions normales : il est plus plat et on observe des zéros à 1,5 et 4 kHz. En revanche, aucun bruit de friction n'est observé. Ces résultats, dans leur ensemble, sont cohérents avec l'hypothèse d'une insuffisante mobilité de la langue, émise à partir de l'analyse des productions vocaliques. En effet, c'est lorsqu'un [i] précède ou suit la plosive [t], que la production spectrale est quasi normale. Or c'est justement dans ce cas que l'amplitude globale du mouvement de la langue est la plus faible. Le long bruit de friction observé dans ce contexte peut, lui aussi, être imputé à une trop grande lenteur de la langue qui, dans ce mouvement de faible amplitude, reste trop longtemps très proche du palais.

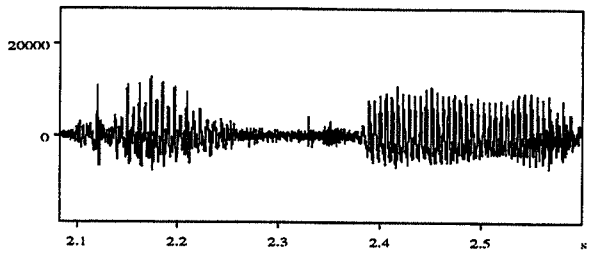
Cette hypothèse est confirmée aussi par les mesures de la tenue consonantique qui, pour [ati] et [ata], est deux fois plus longue que pour un locuteur normal (300 *vs.* 150 ms) et est significativement plus longue pour [ita] (220 *vs.* 150 ms). Les durées du VOT (30 ms) ne dépendent pas du contexte alors que pour les locuteurs normaux le VOT est plus court dans [ata] et [ita] que dans [ati] (30 *vs.* 70 ms). Or, il est admis généralement que, pour les locuteurs normaux, cette différence serait due au dévoisement de la voyelle [i] en contexte gauche [t] [Nea98]. Si on ne retrouve pas cet effet de coarticulation chez le patient, c'est probablement un signe que la stratégie globale de coordination glotte – langue a été altérée.

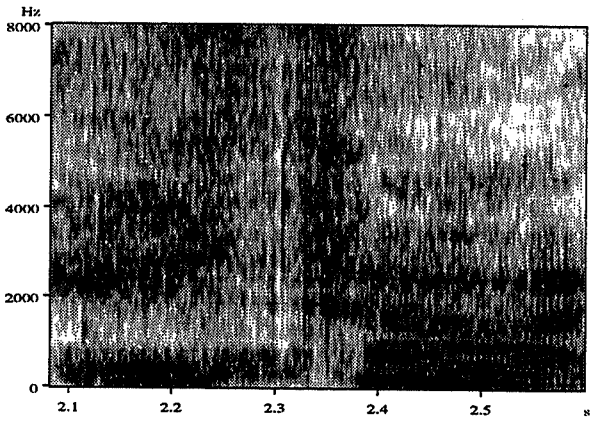
Le [k] du patient #5 est très différent de celui des locuteurs normaux. Le spectre du burst est quasi plat dans les 3 contextes, très similaire à celui du [t] de [ati] pour ce même patient, alors que, pour nos locuteurs de référence, on observe des maxima vers 4 et 7 kHz ainsi qu'un zéro entre ces 2 pics. Ce burst est suivi d'un bruit de friction intense et très long (plus de 100 ms), qui a de l'énergie autour 0,5 kHz et 2 maxima principaux vers 2 et 5 kHz. Les tenues consonantiques sont significativement supérieures à celles des sujets normaux. Cette durée est plus courte pour [ika] que pour [aka] et [aki] (300 *vs.* 350 ms). Le VOT de [ika] et [aka] est 2 fois plus long que pour les sujets normaux (85 *vs.* 40 ms), mais il est 2 fois plus court pour [aki] (27 *vs.* 75 ms). En contexte droit [a], le VOT est plus long pour [k] que pour [t] alors qu'en contexte droit [i], les VOT sont similaires pour les 2 consonnes. Cette différence est en accord avec les données issues de la littérature [Nea98].

À partir de ces mesures, il est possible de déduire que l'hémi-glossectomie subie par le patient #5 génère une forte gêne pour produire des consonnes articulées dans la région vélaire. Au contraire, l'articulation dans la région alvéo-dentale est possible, et nos observations laissent même penser que le [t] et le [k] sont tous les deux produits dans cette région. Malgré cette unicité du lieu d'articulation, le patient #5 préserve la distinction entre ces 2 consonnes, en générant pour le [k] un bruit de friction post-burst plus intense et un VOT plus long.

Pour le patient #7, le spectre de l'explosion du [t] varie selon le contexte : pour [ata], on note 3 maxima vers 3, 5 et 7 kHz ; pour [ati], l'allure générale du spectre est en cloche avec un maximum vers 3,5 kHz ; pour [ita], 2 maxima sont visibles vers 3 et 6 kHz. On observe les mêmes patrons temporels pour les 3 séquences : l'explosion est précédée d'un bruit de friction quasi-périodique, et elle est suivie d'un bruit très énergétique, mélange de friction et de plosion (figures 4A et 4B). Les mesures dans le domaine temporel sont similaires à celles des locuteurs normaux. Pour [k], le spectre du burst est similaire à celui des locuteurs normaux, mais le patron temporel est différent et très comparable à celui du [t], avec les mêmes types de bruit avant et après le burst. Les tenues consonantiques et les VOT sont comparables à ceux des locuteurs normaux.

Ces résultats suggèrent que ce patient arrive à articuler les consonnes au bon endroit, mais éprouve des difficultés pour déplacer sa langue avec les bonnes exigences temporelles, avant et après l'explosion.





Figures 4A et 4B : Signal et sonagramme produit par le patient #7 durant la séquence [ita].

Les résultats obtenus pour le patient #2 sont similaires à ceux des locuteurs normaux aussi bien dans le domaine temporel que spectral. Cependant, en contexte [ita] et [ata] les valeurs de tenue consonantique sont plus longues que les normales (200 *vs.* 150 ms). Ce dernier point semble suggérer que les mouvements articulatoires produits sont plus lents que pour un sujet normal.

5. CONCLUSIONS

L'analyse acoustique des productions vocaliques et consonantiques montre que la parole produite est de qualité différente pour les 3 patients analysés. Cette différence semble en rapport avec l'importance des modifications anatomiques qu'ont causées les interventions chirurgicales. La procédure proposée, fondée sur un corpus composé de logatomes sans signification linguistique et utilisant une analyse classique du signal dans les domaines temporel et fréquentiel, s'avère donc pertinente pour évaluer le degré de gêne des patients. On peut donc envisager de l'utiliser pour des évaluations de l'impact d'interventions chirurgicales d'importances similaires, mais faisant appel à des techniques différentes. Il nous a été aussi possible d'affiner les conséquences des perturbations anatomiques sur les gestes articulatoires utilisés lors de la production de la parole.

L'étude de locuteurs pathologiques, dans la lignée des travaux sur les perturbations artificielles de la parole [Gay81, Sav95], permet aussi d'avoir une meilleure idée de la représentation qu'a un locuteur de la tâche parole. Les conclusions que nous pouvons tirer des résultats exposés ci-dessus confirment celles de nos travaux sur les tubes labiaux : la représentation est clairement dans un espace perceptuel [Sav99] puisque le patient #5, empêché par sa pathologie de produire 2 lieux d'articulation différents pour [t] et [k], maintient une distinction par le biais d'un bruit de friction plus intense et d'un VOT plus long pour le [k]. Mais elle possède probablement des ancrages articulatoires [Sav95], puisque ce même patient respecte la hiérarchie dans le degré d'aperture des voyelles en dépit de l'absence de conséquence acoustique significative.

REMERCIEMENTS

Ces travaux ont été soutenus par une bourse de *La Ligue contre le Cancer.*

BIBLIOGRAPHIE

[Del97] Deleyiannis, FWB., Weymuller, Jr A. and Coltrera MD. (1997), "Quality of life of disease-free survivors of advanced (stage III or IV) oral pharyngeal cancer", *Head & Neck*, 19, 466-473.

[Gay81] Gay T., Lindblom B. and Lubker J. (1981), "Production of bite-block vowels: Acoustic equivalence by selective compensation", *J. Acoust. Soc. Am.*, 69, 802-810.

[Nea98] Neagu A. (1998), "Représentations phonétiques et identification des syllabes occlusive-voyelle en français", *Thèse de l'Institut National Polytechnique de Grenoble*, France.

[Sav95] Savariaux C., Perrier P. and Orliaguet J.P. (1995), "Compensation strategies for the perturbation of the rounded vowel [u] using lip-tube: A study of the control space in speech production", *J. Acoust. Soc. Am.*, 98, 2428-2442.

[Sav99] Savariaux C., Perrier P., Orliaguet J.P. and Schwartz J.L. (1999), "Compensation strategies for the perturbation of French [u] using a lip tube: II. Perceptual Analysis", *J. Acoust. Soc. Am.*, 106 , 381-393.

Parole hyper-articulée : données et analyses acoustiques pour des plosives en français

Denis Beautemps

Institut de la Communication Parlée,
CNRS UMR 5009, INPG/Univ Stendhal,
46 Avenue Félix Viallet,
38031 Grenoble Cedex 1, France
Tél.: ++33 (0)476 57 47 15 - Fax: ++33 (0)476 57 47 10
Mél: beautemps@icp.inpg.fr

ABSTRACT

In the field of speech adaptability to environmental conditions, the gain in intelligibility of hyper auditory speech versus normal speech is investigated. The " Lombard " reflex was used to obtain auditory sequences produced with vocal effort for a set of French /b, d, g, p, t, k/ plosives. An acoustic analysis performed in the burst vicinity shows that the spectral tilt is an important cue to control the hyper – articulation vocal effort.

1. INTRODUCTION : CONTEXTE DE PAROLE HYPER-HYPO

Cette étude sur les plosives du français s'inscrit dans un cadre général de recherche des mécanismes d'adaptabilité en parole, sous les aspects acoustiques, audio - visuels et articulatoires. Ces signaux de la communication parlée sont variables. Cette variabilité n'est pas pur bruit mais gestion contrôlée d'une négociation entre locuteur et auditeur. Ce contrôle actif d'une source naturelle de variabilité est une des manifestations les plus claires de ce que Lindblom ([Lin86]) appelle la nature adaptative de la parole exploitant conjointement l'information mise dans le signal par l'effort du locuteur, et l'information contextuelle qui n'est pas dans le signal acoustique et qui est récupérée par l'auditeur par sa compréhension générale de la situation de communication. C'est le jugement qu'il a de cet équilibre entre ces deux sources d'information qui conduit dans certains cas à hyper-articuler pour mettre plus d'information dans le signal, et dans d'autres cas à «hypo-articuler» et exploiter la quantité d'information importante hors signal afin de minimiser ses efforts d'articulation. Dans ce cadre, cette étude vise à mettre en évidence les stratégies acoustiques de contrôle de l'hyper - articulation à partir de signaux produits en condition de parole hyper-articulée, celle-ci étant naturellement obtenue en situation d'interaction locuteur-auditeur et par réflexe Lombard du locuteur (voir par exemple [Sum88], [Jun96]).

2. DONNÉES

2.1 Corpus et protocole d'enregistrement

Un sujet français prononçant un ensemble de 28 séquences /VCVsV/ constitué des voyelles /a, i, u, y/ et des plosives /b, d, g, p, t, k/ a été enregistré sous deux conditions de bruit environnemental: bruit blanc d'un niveau de 80 dB SPL (étalonné à l'aide d'une oreille artificielle) présenté aux oreilles du locuteur, et sans bruit environnemental ce qui constitue la condition de référence. La condition de bruit environnemental a été introduite pour inciter le locuteur à réaliser un effort vocal naturel. Le locuteur PB avait donc pour tâche de prononcer la phrase «T'as dit /VCVsV/» en présence de l'auditeur qui faisant une erreur sur la consonne, l'incitait à répéter en insistant sur la consonne plosive pour les deux conditions d'environnement. Cette procédure appliquée trois fois a permis de constituer une base de données de 288 stimuli acoustiques appartenant à quatre catégories de production: parole en condition normale, normale répétée, lombard et lombard répétée.

2.2 Données perceptives

Une analyse du gain en intelligibilité de la parole hyper-articulée a été menée à l'aide d'un test d'identification des plosives consistant à mélanger les stimuli originaux avec du bruit blanc b(t) à différents rapports signal sur bruit, pour une réalisation des conditions de parole "normale" et "lombard répétée" ([Bea99]). 2 x 24 stimuli x(t) normalisés en énergie ont été dégradés avec 7 niveaux t de rapport signal sur bruit et présentés à l'oreille gauche de 10 sujets français ne présentant pas de trouble connu de l'audition:

$y(t) = x(t) + \alpha\ b(t)$ avec $\alpha = 10^{-t/20}$ avec b(t) et x(t) d'énergie moyenne identique.

L'analyse des taux d'identification révèle une dégradation du score global lorsque le rapport signal sur bruit décroît (figure 1a) pour les deux conditions de production testées, mais avec une meilleure robustesse de la condition lombard répétée: il est à noter un taux d'identification encore raisonnable de 75 % à t = 0 dB, pratiquement identique au taux en parole « normale » pour t = 12 dB.

L'analyse par consonne ([Bea99]) montre que le gain de la parole lombard est nette pour les lieux d'articulation vélaires et alvéolaires et que le trait de voisement est moins robuste au bruit que le lieu d'articulation. Enfin, le cas des consonnes bilabiales montre (Figures 1b et 1c) que l'effort vocal n'est pas toujours payant en terme d'amélioration de l'intelligibilité. Ceci est vraisemblablement dû au fait que le lieu d'articulation particulier des lèvres ne permet pas de rendre audible un effort vocal articulatori-acoustique.

3. CARACTÉRISATION ACOUSTIQUE

3.1 Corpus analysé

Les données perceptives montrent que dans une situation où un objectif d'identification est demandé, le locuteur sait renforcer de manière sélective certaines phases acoustiques. Cette partie présente donc une étude des stratégies mises en œuvre au cours de l'effort vocal en s'appuyant sur l'analyse acoustique du sous-ensemble composé des plosives non voisées /p/,/t/ et /k/ du corpus pour les quatre conditions de production. Pour les 144 stimuli $/V_1CV_2sV_3/$, les évènements de début de voisement (t_{v1_onset}, t_{v2_onset}, t_{v3_onset}), de centre du noyau vocalique (t_{v1}, t_{v2}, t_{v3}) et de disparition de la structure formantique (t_{v1_end}, t_{v2_end}, t_{v3_end}), ont été repérés pour les trois voyelles ainsi que l'instant (t_{burst_onset}) d'apparition de la barre d'explosion du burst de la consonne plosive.

3.2 Analyse temporelle

Une première analyse a consisté à comparer l'énergie au niveau de chacun de ces évènements (table 1). Pour les voyelles, l'énergie est moyennée sur une durée de 20 ms centrée sur les instants t_{v1}, t_{v2} et t_{v3} tandis que l'énergie localisée sur le burst est calculée sur la durée (t_{v2_onset} - t_{burst_onset}). On note une augmentation générale de l'énergie moyenne dans la condition «lombard répétée» avec notamment un renforcement significatif de la syllabe $/CV_2/$ et un poids plus important du burst de la plosive par rapport à V_2. La figure 2 montre l'effet gradué des quatre conditions de parole sur l'énergie du burst. La durée du burst (figure 3) ne semble pas sur ces données être influencée par la condition de production. La durée de l'occlusion (figure 4) (mesurée entre t_{v1_end} et t_{burst_onset} [Abr90]) est pour sa part sensible à l'effet d'insistance sur la consonne (parole répétée et lombard répétée confondues) pour lequel il est noté une augmentation moyenne de 75 ms par rapport à la condition normale (parole normale et parole lombard confondues). On remarque également que l'effet conjoint d'insistance et de réflexe lombard a pour conséquence l'amplification de ce phénomène, avec un gain moyen significatif de 35 ms entre la condition de parole répétée et de parole lombard répétée.

Ces résultats montrent qu'au delà d'un effet «volumique» d'augmentation de l'énergie, l'hyper - articulation sur la consonne se caractérise par un renforcement dans la région du burst que nous proposons d'examiner par une analyse spectrale.

Table 1: Energies en dB (Moyenne | Ecart-type) au niveau de V_1, V_2, V_3 et du burst de la plosive C pour les séquences $/V_1CV_2sV_3/$ produites sous les conditions «normale» et «lombard répété».

	Normale	Lombard répété
V_1	56,41 | 2,05	59,03 | 3,23
Burst	38,52 | 4,18	49,30 | 4,33
V_2	55,84 | 2,79	64,26 | 3,43
V_3	48,16 | 1,67	60,19 | 1,96

3.3 Propriétés du spectre de la plosive entre t_{burst_onset} et t_{v2_onset}

Les signaux acoustiques de la région du burst (entre t_{burst_onset} et t_{v2_onset}) d'énergie moyenne normalisée à un sont préalablement décomposés en bandes spectrales suivant un banc de 8 filtres de gain unité [Tes99] répartis régulièrement entre 0 et 21.3 Bark (8000 Hz) (figure 5).

Une analyse en composantes principales de l'énergie obtenue en sortie du banc de filtres fait émerger pour les signaux de la condition de production «normale» deux facteurs expliquant 72 % de la variance globale. La projection de l'ensemble des données suivant les deux axes principaux montre pour la condition «lombard répétée» une discrimination améliorée entre /p/ et /t/ le long du premier axe Vp_1 et suivant le second Vp_2, une meilleure distinction des catégories /p/, /t/ et /k/ (figure 6) alors que seule la catégorie /t/ émerge le long de Vp_1 pour la condition de parole lombard (figure 7). Le premier facteur principal P_1 présente une forte corrélation (r = 0,9) avec la pente en dB par octave du spectre du burst: pente_db_par_octave = 0,9656 . P_1 – 3,1742. La contribution à la valeur du second facteur (table 2) est essentiellement due aux bandes 4 à 6 du banc de filtres (fréquences centrales entre 1150 et 2700 Hz), correspondant à un indice d'énergie en moyenne fréquence.

Ces premiers résultats montrent la tendance en "lombard répété" à une baisse de la pente spectrale pour privilégier les basses fréquences pour /p/, à favoriser les hautes fréquences par une augmentation de la pente spectrale pour /t/ et à augmenter l'énergie en moyenne fréquence (entre 1150 et 2700 Hz) pour /k/. Ce jeu sur la pente a été également observé par Castellanos et al. [Cas96] lors de l'étude de l'effet Lombard en Espagnol, notamment pour la plosive /d/ en Espagnol. Ces indices de pente montante, pente descendante, et d'énergie en moyenne fréquence sont à rapprocher des indices «diffuse-rising», «diffuse-falling» et «compact» obtenus pour caractériser respectivement les lieux d'articulation alvéolaire, labial et vélaire des consonnes par Blumstein et Stevens [Blu79].

Table 2: Coordonnées des deux premiers vecteurs de l'analyse en composantes principales.

Num. Filtre	Vp_1	Vp_1
1	-0,1688	0,1371
2	-0,4615	-0,1370
3	-0,4514	-0,1451
4	-0,3530	0,3081
5	-0,2222	0,6363
6	0,1534	0,6568
7	0,4306	0,1011
8	0,4145	0,0007

4. CONCLUSION

Un cas de parole hyper – articulée pour des plosives a été mis en évidence sous ces aspects perceptifs et acoustiques à l'aide du réflexe lombard provoqué chez un locuteur français par application d'un bruit blanc de 80 dB présenté à ces oreilles. L'analyse acoustique montre que dans cette situation, le locuteur cherche à contrôler la pente spectrale dans la région d'explosion pour les plosives non voisées labiales et dentales et tente un renforcement dans la zone de moyenne fréquence pour les vélaires. Cette étude nécessite d'être poursuivie par une caractérisation articulatoire de l'effort vocal à l'aide d'outils tel que l'articulographe, pour notamment étudier les réorganisations des gestes articulatoires durant la phase expansive de l'occlusion dans le cas des consonnes bi-labiales. Enfin, l'ensemble de ce travail reste à être étendu à plusieurs locuteurs.

Remerciements : Ce travail a bénéficié de discussions fructueuses et de l'aide de P. Badin, G. Bailly, F. Berthommier, P. Borel, J.L. Schwartz et C. Savariaux.

BIBLIOGRAPHIE

[Abr90] Abry C., Orliaguet J.-P. & Sock R. (1990) « Patterns of sppech pahsing. Their robustness in the production of a timed linguistic task : single vs. double (abutted) consonants in French », Cahiers de Psychologie Cognitive, European Bulletin of Cognitive Psychology, Vol. 10, n°3, pp. 269-288.

[Bea99] Beautemps D., Borel P. & Manolios S. (1999) " Hyper-articulated speech: Auditory and visual intelligibility", in proceedings of the 6th European Conference on Speech Production (Budapest), Vol.1, pp. 109-112.

[Blu79] Blumstein S. E., Stevens K. N. (1979) "Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants", J. Acoust. Soc. Am. 66(4), pp. 1001-1017.

[Cas96] Castellanos A., Benedi J.M., Casacuberta F. (1996) « An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect », Speech Communication, 20, pp. 23-35.

[Jun96] Jean-Claude Junqua (1996) « The Influence of acoustics on speech production : A noise-induced stress phenomenon known as the Lombard reflex », Speech Communication, 20, pp. 13-22.

[Lin86] Lindblom B. (1986-1987) "Adaptative variability and absolute constancy in speech signals : two themes in the quest for phonetic invariance", Perilus, 5, pp. 2-20.

[Sum88] Summers W., Pisoni D., Bernacki R., Pedlow R. & Stokes M. (1988) "Effects of noise on speech production: Acoustic and perceptual analyses", J. Acoust. Soc. Am., 84 (3), pp. 917-928.

[Tes99] Tessier E., Berthommier F., Glotin H.& Choi S. (1999) "A casa front-end using the localisation cue for segregation end then cocktail-party spech recognition ", in proceedings of the ICSP'99, Séoul, pp. 97-102.

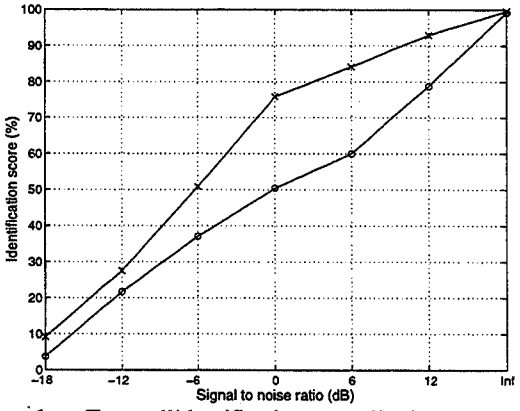


Figure 1a: Taux d'identification en % (toute consonne confondue) en fonction du rapport signal sur bruit. Condition "lombard répétée" (lignes avec croix) et condition normale (lignes cerclées).

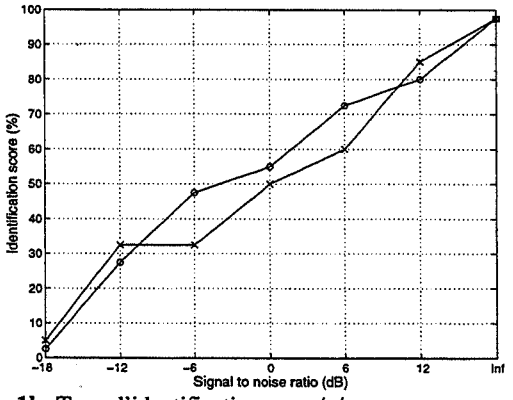


Figure 1b: Taux d'identification pour /p/.

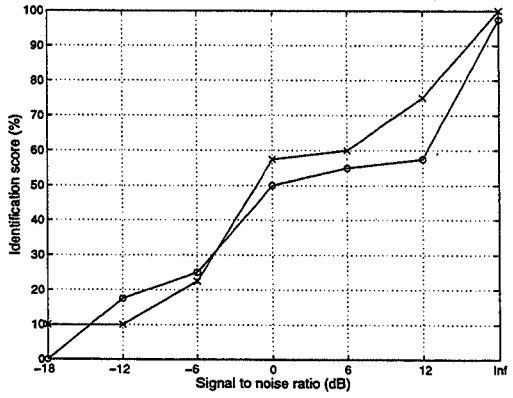


Figure 1c: Taux d'identification pour */b/.*

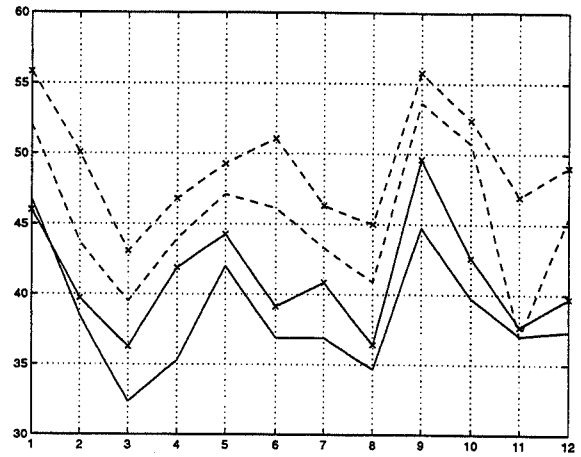


Figure 2: Energie du burst (en dB). La moyenne des trois réalisations des séquences est présentée suivant l'ordre en abscisse: /apa/, /ipi/, /upu/, /ypy/, /ata/, /iti/, /utu/, /yty/, /aka/, /iki/, /uku/, /yky/. En trait plein, condition de parole normale, en tirets, condition Lombard, et représentation en croix pour l'effet d'insistance.

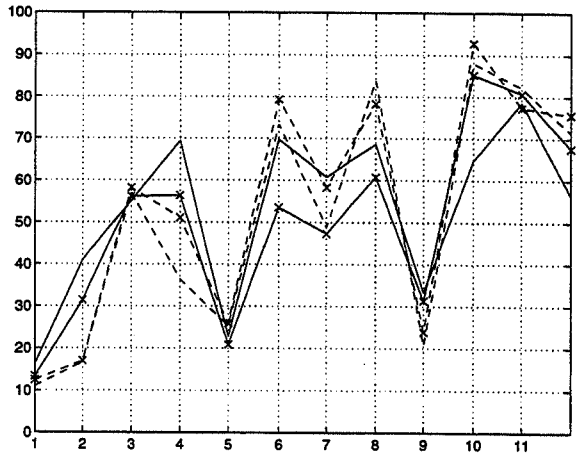


Figure 3: Durée du burst (en ms). La moyenne des trois réalisations des séquences est présentée suivant l'ordre en abscisse: /apa/, /ipi/, /upu/, /ypy/, /ata/, /iti/, /utu/, /yty/, /aka/, /iki/, /uku/, /yky/. En trait plein, condition de parole normale, en tirets, condition Lombard, et représentation en croix pour l'effet d'insistance.

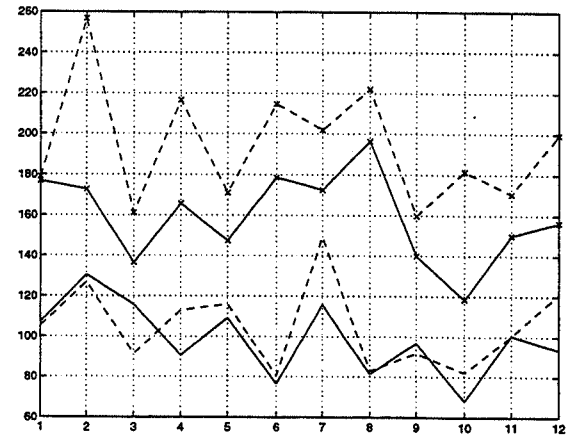


Figure 4: Durée de l'occlusion (en ms). La moyenne des trois réalisations des séquences est présentée suivant l'ordre en abscisse: /apa/, /ipi/, /upu/, /ypy/, /ata/, /iti/, /utu/, /yty/, /aka/, /iki/, /uku/, /yky/. En trait plein, condition de parole normale, en tirets, condition Lombard, et représentation en croix pour l'effet d'insistance.

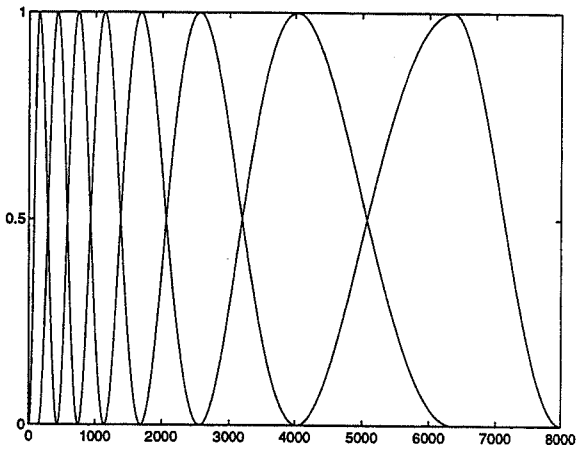


Figure 5: Fonction de transfert des 8 filtres suivant leur position entre 0 et 8000 Hz. Fréquence centrale du premier filtre : 1,33 Bark. Largeur de bande à -3dB : 2,66 Bark.

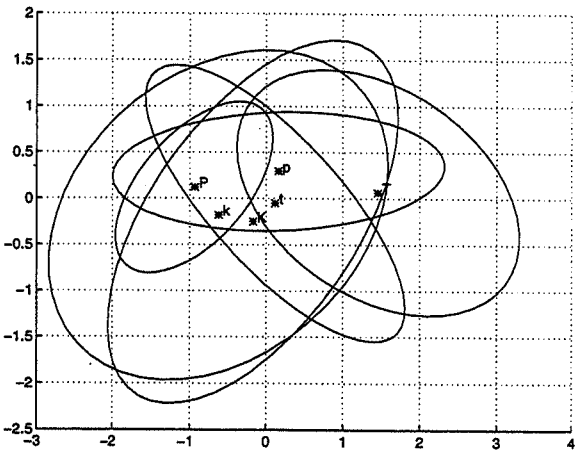


Figure 6: Représentation selon les deux premiers axes principaux (Vp_1, Vp_2) de la moyenne et de la dispersion à 1 écart-type des deux premiers facteurs P_1 et P_2 pour /ptk/ en condition de parole « normale » (minuscule) et « lombard répétée » (majuscule).

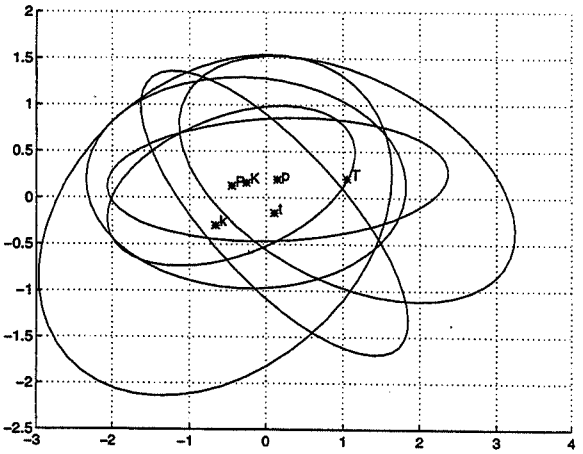


figure 7: Représentation selon les deux premiers axes principaux (Vp_1, Vp_2) de la moyenne et de la dispersion à 1 écart-type des deux premiers facteurs P_1 et P_2 pour /ptk/ en condition de parole « normale » (minuscule) et « lombard » (majuscule).

Evaluation objective de la dysprosodie des pathologies neurologiques: critères de différenciation diagnostique et suivi longitudinal des prises en charge thérapeutiques

Bernard Teston, Alain Ghio, François Viallet

Laboratoire Parole et Langage
ESA 6057 CNRS Université de Provence
29 avenue Robert Schuman - 13621 Aix en Provence France
e-mail teston@lpl.univ-aix.fr

ABSTRACT

We present a diagnosis aid and therapeutic follow-up method for neurological pathologies in the context of a clinical research project. This method is based on the analysis of prosodic dysfunctionings caused by illness. It take into account the three prosodic parameters : pitch, intensity and duration.

Initial results indicated that the method holds its promise. We are in the process of constructing a knowledge base of numerous patients assessed with this method to enhance its performance.

1. INTRODUCTION

Les troubles de la communication verbale, provoqués par des dysfonctionnements neuro pathologiques, représentent un facteur de handicap social de plus en plus déterminant dans l'environnement socio-économique actuel où l'autonomie de chaque individu dépend d'un nombre considérable d'interactions. La conséquence la plus importante de ces troubles se situe dans l'intelligibilité de la parole, souvent dans des conditions de communication nouvelles. Parmi ces troubles, les dysprosodies, bien qu'elles affectent la production vocale, n'ont jamais été traitées en tant que telles, leurs améliorations étant liées à l'état d'évolution de la maladie.

Il apparaît depuis peu, que certains paramètres prosodiques peuvent donner des indications quantitatives pertinentes sur cet état . Nous avons dans ce but déposé et obtenu un projet de recherche clinique (PHRC) présenté par le Centre Hospitalier régional d'Aix, avec pour ambition de développer une évaluation instrumentale multiparamétrique des dysprosodies par atteinte neuro-motrice en vue d'améliorer leur prise en charge diagnostique et thérapeutique, incluant la réadaptation fonctionnelle.

Cette investigation doit, au terme du projet, proposer une méthode basée sur l'enregistrement du seul signal de parole microphonique associé à des mesures physiques et statistiques des données cliniques sur micro ordinateurs de type PC. D'une utilisation de routine, cette méthode est non invasive et aisée à mettre en oeuvre en pratique ambulatoire dans la prise en charge des patients (consultation en hôpital de jour). Ce sont les premiers résultats de cette action en terme de développement méthodologique et de base de connaissances qui sont exposés dans la présente étude.

2. LES DYSARTHRIES

On nomme "dysarthrie" l'ensemble des troubles de la parole liés à des perturbations des commandes neuro-musculaires des organes mis en jeux dans la production de la parole, dont l'origine est une lésion du système nerveux central ou périphérique. Le terme de dysarthrie comprend donc tous les dysfonctionnements relatifs à la respiration, phonation, articulation et prosodie. Les dysarthries sont schématiquement caractéristiques de certains symptômes associés aux grandes familles d'affections neurologiques ayant une influence plus ou moins grande sur la production de la parole : rigidité, incoordination, paralysie, et spasmes. Il est donc légitime de chercher à les analyser objectivement dans le but de permettre une aide au diagnostic ainsi qu'un suivi thérapeutique longitudinal. Ceci d'autant plus, que l'acte de parole est au plan des programmations et coordinations neuro-motrice plus complexe que la marche ou la préhension, traditionnellement décrites comme évaluateurs des niveaux de pathologie. On reconnaît schématiquement trois grandes classes de dysarthries :
Les dysarthries hypokynétiques qui se caractérisent par une réduction de la dynamique des mouvements [Gent95].

Les dysarthries ataxiques qui provoquent la perte de la coordination des mouvements.

Les dysarthries paralytiques qui sont des paralysies provoquées par l'atteinte dégénératives des motoneurones.

3. L'EVALUTION DES DYSARTHRIES

L'évaluation des dysarthries se fait à l'écoute du patient en portant son attention sur les aspects caractéristiques de la production de sa parole. On s'attache particulièrement à la netteté et la précision des voyelles et des consonnes, à la réalisation de groupe de consonnes (coarticulation). La prosodie est étudiée à travers la facilité du discours, la longueur des pauses, les changements de rythme et dans son maintien au niveau mélodique ou accentuel. La voix est surtout évaluée au niveau de son souffle et de sa stabilité à moyen terme (tremor). On peut objectiver l'évaluation des dysarthries au moyen de mesures physiques sur les paramètres acoustiques, aérodynamiques et kynésiographiques de la production de la parole.

4. PROBLEMATIQUE GENERALE

4.1 L'évaluation des dysprosodies

Il apparaît de ce rapide descriptif des différents moyens d'évaluer les dysfonctionnements articulatoires que les paramètres prosodiques, mélodie, accent et débit vocal (rythme pause et débit) présentent un bon compromis entre leur facilité de capture, de mesure de représentation, et leur correspondance à l'état pathologique.

En pratique clinique, l'ensemble des échelles d'évaluation accorde dans la parole une place croissante à l'analyse des dysprosodies. Mais la description pragmatique reste globale et qualitative en l'absence de variables instrumentales plus élémentaires et validées, se prêtant mieux à une analyse quantitative.

La prise en charge actuelle des affections neurologiques nécessite une évaluation diagnostique précise conduisant à catégoriser les patients dans des classifications nosologiques, dont les frontières, souvent mal définies, évoluent avec la progression des connaissances en neurobiologie. A l'analyse clinique traditionnelle sont ainsi venus s'ajouter divers critères de classification de nature instrumentale, basés sur les données de l'imagerie cérébrale fonctionnelle et anatomique, de la pharmacologie ou encore de la biologie moléculaire et de la génétique.

Certains processus neurologiques, comme la maladie de Parkinson ou l'ensemble élargi des syndromes parkinsoniens qui comportent dans leur expression clinique commune une dysprosodie, restent encore de bons candidats à une analyse instrumentale du symptôme en question, en vue d'une classification plus discriminante que celles obtenues sur la base du seul critère clinique en l'absence d'informations pertinentes disponibles à partir des autres critères instrumentaux (imagerie, biologie moléculaire, données neuro-pathologiques).

Les approches thérapeutiques à visée symptomatique nécessitent aussi, pour être évaluées objectivement, une représentation quantifiée du symptôme-cible: il en est ainsi, dans le contexte des troubles de la production vocale, par exemple, pour la prise en charge en rééducation des dysprosodies parkinsoniennes ou pour le traitement par injection de toxine botulinique des dystonies laryngées et oro-mandibulaires.

Un grand nombre d'études descriptives ont été consacrées à l'analyse des dysprosodies au cours de diverses affections du système nerveux central (Kent et Rosenbek [Ken82]). Elles ont surtout porté sur les dysprosodies hypokynétiques. Les études sur la mélodie ont montré des résultats souvent contradictoires, en particulier dans la maladie de Parkinson. La Fo augmente avec la sévérité du trouble chez Herlich et Ackerman [Her93], et Ludlow et Bassich [Lud83].

Elle diminue par contre chez Canter [Can63]] et de nombreux autres auteurs. D'une manière générale, on observe une plage de variation de la Fo nettement plus réduite responsable d'une parole monotone (Weismer [Wei84]). La vitesse d'élocution varie également dans de grande proportion. Elle est très ralentie avec des pauses longues pour les dysarthries ataxiques. Chez les parkinsoniens en revanche elle est caractérisée par une accélération du débit avec des anomalies de la segmentation rythmique (Darkins et al [Dar88], Caekebeke et al, [Cae91), mais ces résultats sont très variables, Volkman et al [Vol92], trouvent même une lenteur de parole chez tous les patients parkinsoniens.
Toutes ces études estimables sont cependant entachées par de nombreux travers. Elle sont d'une manière générale très parcellaires, établies sur des populations souvent restreintes aux symptômes cliniques parfois mal définis. Elles manquent d'homogénéité dans le choix des paramètres pertinents et surtout des méthodes d'évaluation qui sont rarement bien maîtrisées.

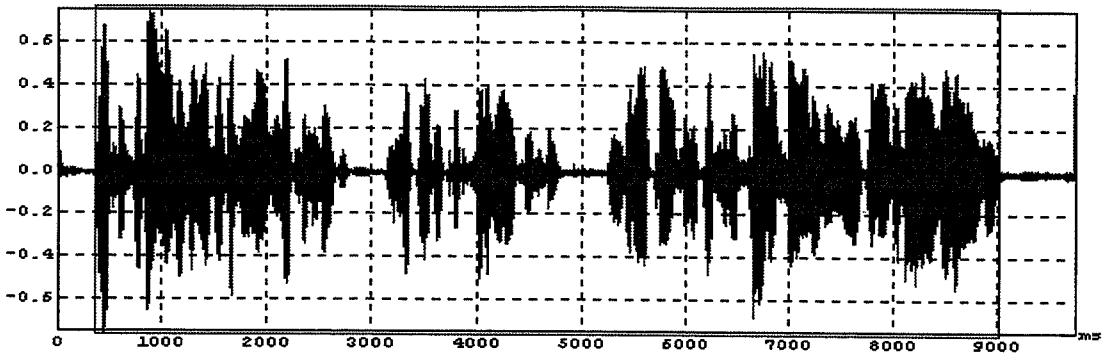
4. 2 Méthodologie

Dans le souci de pouvoir mener les évaluations de la dysprosodie de la parole de la manière la plus aisée et la plus conviviale possible par un personnel médical ou para médical non spécialiste nous les avons intégrées sous la forme de programmes dédiés à la station EVA d'aide au diagnostic et à la rééducation des pathologies de la voix et de la parole (Teston et Galindo [Tes95]). Ce matériel d'investigation clinique dans les domaines de l'ORL et la neurologie, maintenant bien stabilisé après plusieurs années de mise au point et de tests d'utilisation est devenu un standard dans l'évaluation des dysphonies. DIANA en est un dérivé plus simple basé exclusivement sur l'analyse du signal de parole. Il fonctionne sous la forme de station de travail sur PC dans l'environnement WINDOWS dont les différentes fonctions se présentent sous la forme d'applications dédiées à un problème clinique particulier.

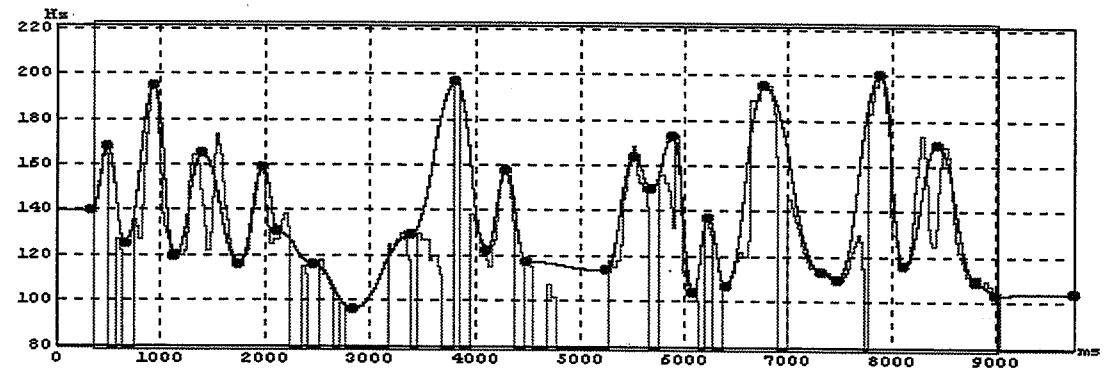
L'application à l'étude de la dysprosodie est basée sur la mesure et l'interprétation des trois paramètre prosodiques ; mélodie, intensité et durée sur des phrases types lues par les patients ou répétées selon un modèle. Il est possible d'utiliser de la parole spontanée mais les résultats étant très dépendants du support d'élocution il est conseillé d'en conserver le même type pour des études comparatives. La durée moyenne des phrases lues est de l'ordre d'une minute pour une évaluation pertinente. Sur ces supports d'élocution sont calculées des représentations graphiques et statistiques sous forme de graphes aisément interprétables ainsi que des indices chiffrés : moyenne , mode, écart type, coefficient de variation, dynamique Min-Max.

La Fo est calculée au moyen d'une méthode AMDF sur une durée de 30 ms et un pas de 10 ms, après détection du voisement par filtrage passe bas et détection des passages par zéro. Cette méthode a été choisie pour sa bonne robustesse au timbre souvent dégradé des voix pathologiques et sa précision. La courbe de Fo obtenue est modélisée au moyen de la méthode MOMEL (Hirst et Espesser [Hir93]). Elle est basée sur la détection de

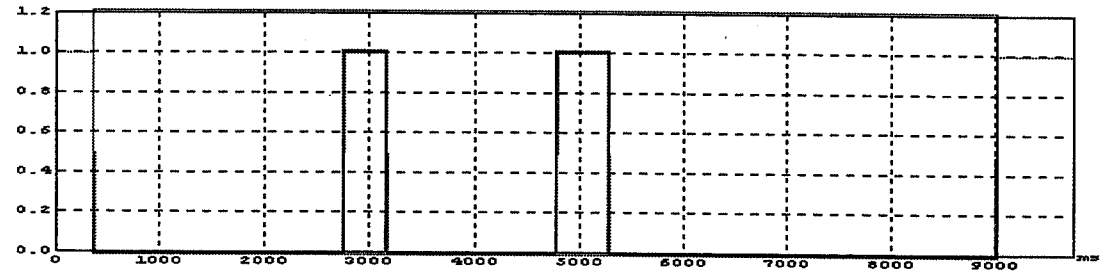
Oscillogramme



Fréquence fondamentale modélisée



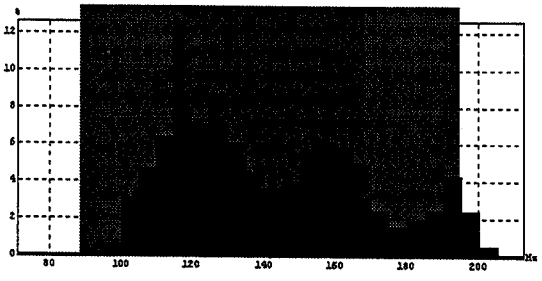
Détection des pauses silencieuses



Statistiques F0

	Fréquence	½tons
Moyenne	141.7 Hz	Do#2
Mode	[115.0 120.0] Hz	Sib1
Ecart-type	26.6 Hz	3.0
Coeff. de variation	18.8 %	-
Min	97.6 Hz	Sol1
Max	200.7 Hz	Sol2
Dynamique	103.1 Hz	12.5

Histogramme F0



Statistiques des pauses silencieuses

	Pause	Signal	Total
Temps cumulé (sec.)	0.9000	7.750	8.650
Répartition	10.4 %	89.6 %	100 %
Nb.	2	3	5
Durée moyenne (sec.)	0.450	2.583	1.730

Exemple d'analyse prosodique sur l'énoncé :" M.Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s'en allaient dans la montagne et là-haut le loup les mangeait... ". ***D'autres exemples peuvent être consultés sur le CD-ROM.***

points cibles qui correspondent aux variations locales pertinentes de la courbe mélodique, reliés par une interpolation de type spline quadratique. Ses variations correspondent alors au profil suprasegmental de la phrase dans lequel n'interviennent pas les variations micromélodiques et les interruptions dues aux dévoisements. Ce profil représente donc la courbe de programmation mélodique de la phrase. Les analyses statistiques sont réalisées sur les données modélisées, les parties présentes dans les pauses silencieuses en sont exclues. Deux distributions sont représentées : celle des valeurs de Fo et celle des valeurs de variation entre les points cibles.

La courbe d'intensité est représentée par la valeur efficace de la pression acoustique en dB avec un pas de 10 ms. La encore elle est modélisée sous la forme de points cibles détectés sur les min et max d'énergie, les traitements statistiques étant identiques à ceux de la mélodie.

Le paramètre de la durée est basé sur le traitement des pauses à partir de l'énergie de la pression acoustique des zones silencieuses ou bruitées du signal. Seules les pauses silencieuses sont identifiées. La distribution des pauses de la phrase est présentée avec les mêmes caractéristiques que précédemment. Les données numériques fournissent le temps cumulé des pauses, du signal, leur répartition, leur nombre et moyenne.

5. CLASSIFICATION CLINIQUE ET CHOIX DES CRITERES

Les informations données par les paramètres prosodiques de cette application sont en voie de validations sur un grand nombre de patients dans le cadre d'un contrat de recherche clinique (PHRC) du ministère de la santé. Nous avons déjà mené une étude sur l'évaluation de l'efficacité des traitements par la L-Dopa sur des patients parkinsonniens (Lagrue et al [Lag99], Meynadier et al [Mey99]). Une étude est en cours sur l'influence prosodique de la stimulation sous thalamiques dans la même affection.

Les trois paramètres prosodiques ont souvent, dans les maladies neurologiques, des perturbations corrélées. Cependant, pour le suivi thérapeutique, la dynamique vocale ainsi que le déplacement de la distribution de la Fo semblent donner dans l'état actuel de nos investigations les informations les plus pertinentes. La modélisation des points cibles mélodiques est également performante pour l'évaluation de l'état ou de l'évolution des dysarthries hypokinétiques. Nous étudions actuellement ses performances comparées avec la modélisation de l'intensité. Cette dernière est par contre très utile pour l'évaluation des tremblements de diverses origines (Tremor). Enfin, la distribution des pauses semble être un bon indice d'évaluation des dysarthries ataxiques. La combinaison des trois paramètres doit nous permettre de distinguer des états cliniques proches mais d'origines différentes ainsi que l'évaluation de l'efficacité thérapeutique des divers traitements proposés. Cependant, cela ne pourra être possible qu'après l'application de connaissances que nous ne maîtrisons pas encore. Ceci est l'objectif da la base que nous sommes en train de constituer au moyen de la méthode que nous venons de décrire et qui contient déjà trois cents patients.

BIBLIOGRAPHIE

[Cae91] Caekebeke, J.F.V., Jenneken-Schinkel, A., Van Der Linden, M.E., Buruma, A.J.S. & Roos, R.A.S. (1991). The interpretation of dysprosody in patients with Parkinson's disease, *J. Neurol. Neurosurg. Psychiatry*, Vol. 54, pp 145-148.

[Can63] Canter, G. J. (1963). Speech characteristics of patients with Parkinson's disease: Intensity, pitch and duration, *J. Speech Hearing Des.*, Vol 28, pp 217-224.

[Dar88] Darkins, A.W. & Fromkins, V.A., Benson, D.F. (1988). A characterization of the prosodic loss in Parkinson's disease, *Brain Lang.*, Vol. 34. pp 317-327.

[Gen95] Gentil, M., Pollak, P. & Perret, J. (1995). La dysarthrie parkinsonnienne, Rev. Neurol., vol 151, n°2, pp 105-112.

[Hir93] Hirst, D. & Espesser, R. (1993).Automatic modelling of fundamental frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix*, vol 15, 71-85.

[Ken82] Kent, R. D. & Rosenbek, J.C.(1982). Prosodic disturbance, *Brain Lang.*, Vol. 15, pp 259-291.

[Lag99] Lagrue, B., Mignard , P., Viallet, F. & Gantcheva, R. (1999). Voice and Parkinson disease : A study of pitch, tonal range and fundamental frequency variations, *ICPhS 99 San Francisco*, vol 9, pp 1811-1814.

[Lud99] Ludlow, C.L., Connor, N.P. & Bassich, C.J. (1987). Speech timing in Parkinson and Huntington's disease. *Brain Lang.*, Vol. 32, pp 195-214.

[Mey99] Meynadier, Y., Lagrue, B., Mignard P. & Viallet, F. (1999). Effects of L-Dopa treatment on the production and perception of parkinsonian vocal intonation, *Parkinsonism and Related Disorders 5* , Vancouver, S121.

[Tes95] Teston B. & Galindo B., (1995). A diagnostic and rehabilitation aid workstation for speech and voice pathologies, *Eurospeech 4*, European. Speech Communication Association, Madrid, sept. Vol. 95, pp 1883-1886.

[Vol92] Volkman, J., Hefter, H., Lange & H. W., Freund, H. J., (1992).Impairment of temporal organization of speech in basal ganglia diseases, *Brain Lang.*, Vol. 43, pp 386-399.

[Wei84] Weismer, G., (1984). Acoustic description of dysarthric speech : Perception correlates and physiological inferences, *In : Rosenbeck, C. J. (ed), Seminar in speech and language, Thieme Stratton*, New York, p 324.

Une étude EPG de la palatalisation des occlusives vélaires en français

Caroline Corneau, Alain Soquet et Didier Demolin

Laboratoire de Phonologie
Université Libre de Bruxelles, Belgique
Tél.: ++32 (0)2 650 20 18 - Fax: ++32 (0)2 650 20 07
E-mail : ccorneau@ulb.ac.be - http://www.ulb.ac.be/philo/phonolab

ABSTRACT

This paper examines with electropalatography the production of velar plosives of two Belgium French speakers. Data is analysed with respect to two moments (beginning and release of closure), [voice] feature and vocalic environment. This study characterises palatalisation which results from adjacency with palatal vowels according to the place of articulation and to the amount of lateral tongue/palate contact. Coarticulation phenomena are also described in a temporal perspective.

1. INTRODUCTION

La palatalisation des occlusives vélaires du français en contexte de voyelles palatales est un phénomène bien établi (notamment depuis [Rou01]). Cet article caractérise l'articulation des occlusives vélaires du français à l'aide de l'électropalatographie. Un atout majeur de cette technique est qu'elle permet d'étudier l'aspect dynamique des gestes d'occlusion et des interactions entre gestes consonantiques et vocaliques. Cette étude examine donc l'organisation temporelle du geste d'occlusion et de la palatalisation des occlusives vélaires qui résulte de la présence d'une voyelle palatale adjacente.

Cette étude s'insère dans le cadre d'une recherche dans laquelle les données EPG sont confrontées à des données d'IRM pour caractériser l'articulation des occlusives vélaires. En effet, dans certains contextes vocaliques, le lieu d'articulation de ces consonnes est postérieur au palais EPG, qui ne fournit donc qu'une information incomplète sur la configuration de la langue pendant l'occlusion. Cependant, ces informations permettent d'identifier et de caractériser les cas de palatalisation de ces consonnes vélaires.

2. MATÉRIEL ET MÉTHODE

Les données furent acquises via la station de travail Physiologia [Tes90]. L'électropalatographe (EPG) est le système EPG2 de Reading, qui utilise un palais artificiel en acrylique de 1,5 mm d'épaisseur contenant 62 électrodes de contact (6 électrodes sur le rang le plus antérieur, et 8 électrodes sur les 7 autres rangs) [Har89]. Les données EPG sont présentées sous la forme de profils de contact indiquant le nombre d'électrodes contactées (en ordonnée) sur chaque rang du palais (en abscisse). Un nombre inférieur à 8 contacts par rang indique une constriction et renseigne sur l'importance des contacts latéraux de la langue au palais.

Le corpus est constitué de 72 items produits par chacun des 2 locuteurs masculins belges (**S1** et **S2**) dans la phrase " Papa frappe /item/ parfois ". Le corpus est constitué de l'ensemble des combinaisons de séquences **V1CV2** où C représente /k/ ou /g/, et V une des voyelles /a, ɛ, i, ɔ, u, y/. Deux moments sont choisis dans la production des consonnes pour extraire les données EPG : le début de l'occlusion (**t1**) et le relâchement de l'occlusion (**t2**) (d'après le signal et le spectrogramme).

Les résultats sont évalués statistiquement par un modèle MANOVA – Repeated measures, avec comme facteurs 'within-subjects' les 8 rangs du palais EPG et les 2 temps t1 et t2, et comme facteurs 'between-subjects' les 6 V1, les 6 V2 et les 2 consonnes. Les variables dépendantes sont les 8 mesures du nombre de contacts sur chaque rang. Pour les analyses de variance faites à chacun des deux moments t1 et t2, les résultats des deux locuteurs sont évalués séparément. Les tests post-hoc sont basés sur l'indice de Scheffé. Lorsqu'il n'est pas précisé, le taux de signifiance est de $p<.05$.

3. RÉSULTATS

3.1 Comparaison entre /k/ et /g/

La figure 1 compare le profil des occlusives vélaires sourdes et sonores (S1 et S2 groupés) à t1. Les profils de /k/ et /g/ présentent des différences significatives ($p<.005$), le contact langue/palais étant plus important dans le cas de /g/. Cependant, cette différence ne varie pas selon les rangs du palais EPG ($p<.1$), ni au cours du temps ($p=.08$), ni d'après l'environnement vocalique ($p<.5$).

3.2 Comparaison entre t1 et t2

La figure 2 présente les profils de l'occlusive vélaire (S1 et S2 groupés, /k/ et /g/ groupés) à t1 et à t2. Les différences entre t1 et t2 sont significatives ($p<.001$), et reflètent une interaction avec les rangs du palais ($p<.001$). Le relâchement de l'occlusion (t2) présente un profil avec des contacts plus importants sur les rangs postérieurs. Ce résultat suggère que le début et la fin de l'occlusion présentent des différences indépendantes de l'environnement vocalique qui semblent être dues à la temporalité même du geste d'occlusion.

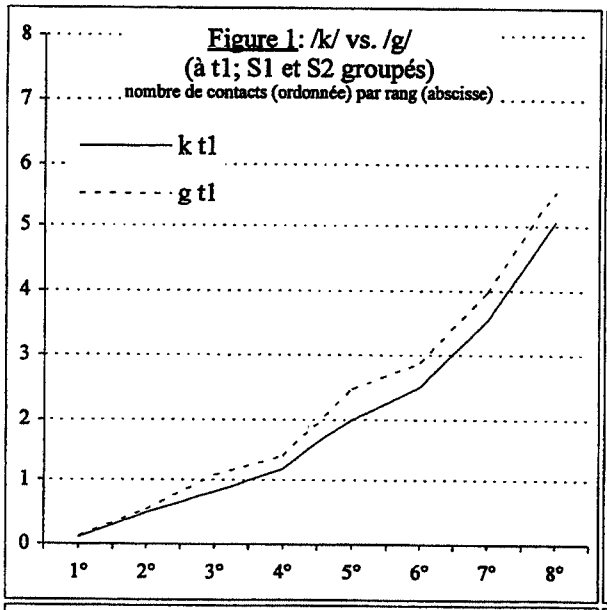


Figure 1: /k/ vs. /g/
(à t1; S1 et S2 groupés)
nombre de contacts (ordonnée) par rang (abscisse)
k t1
g t1
1° 2° 3° 4° 5° 6° 7° 8°

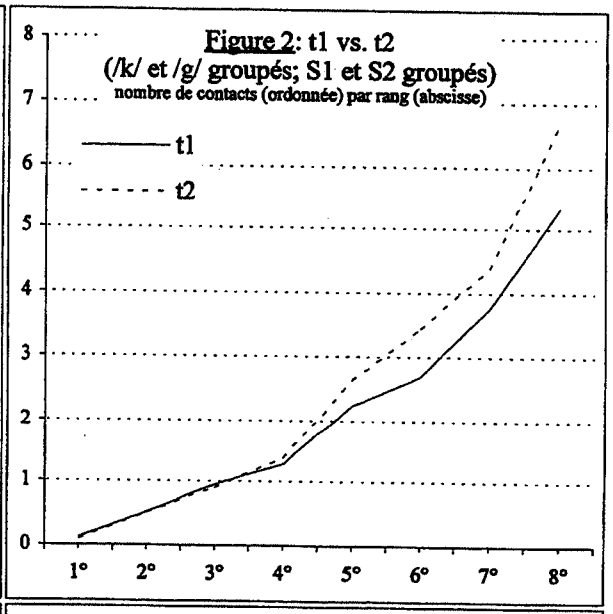


Figure 2: t1 vs. t2
(/k/ et /g/ groupés; S1 et S2 groupés)
nombre de contacts (ordonnée) par rang (abscisse)
t1
t2
1° 2° 3° 4° 5° 6° 7° 8°

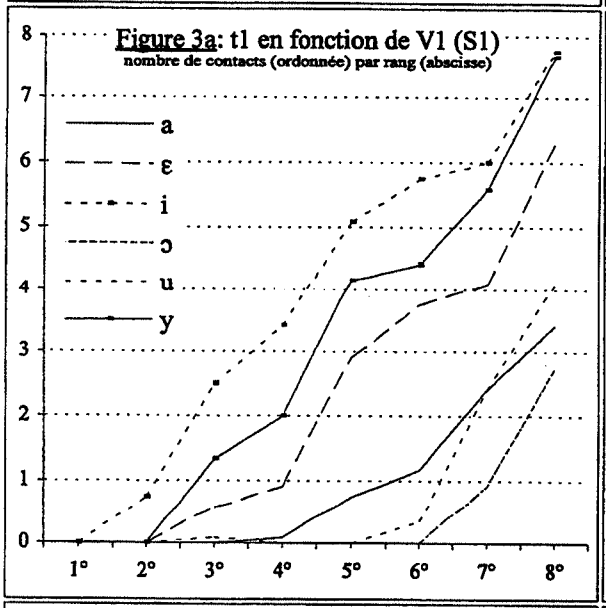


Figure 3a: t1 en fonction de V1 (S1)
nombre de contacts (ordonnée) par rang (abscisse)
a
ɛ
i
ɔ
u
y
1° 2° 3° 4° 5° 6° 7° 8°

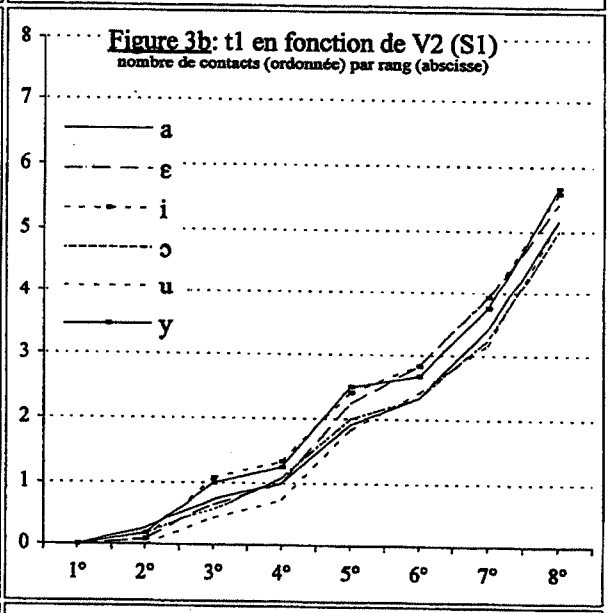


Figure 3b: t1 en fonction de V2 (S1)
nombre de contacts (ordonnée) par rang (abscisse)
a
ɛ
i
ɔ
u
y
1° 2° 3° 4° 5° 6° 7° 8°

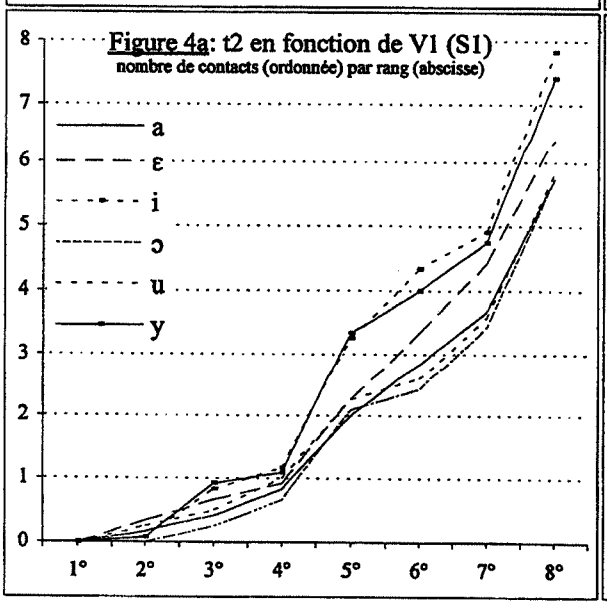


Figure 4a: t2 en fonction de V1 (S1)
nombre de contacts (ordonnée) par rang (abscisse)
a
ɛ
i
ɔ
u
y
1° 2° 3° 4° 5° 6° 7° 8°

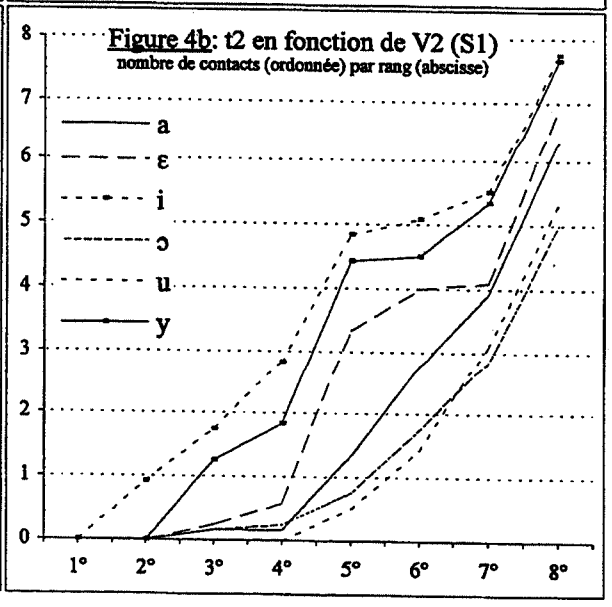


Figure 4b: t2 en fonction de V2 (S1)
nombre de contacts (ordonnée) par rang (abscisse)
a
ɛ
i
ɔ
u
y
1° 2° 3° 4° 5° 6° 7° 8°

3.3 Caractérisation du début de l'occlusion (t1)

En fonction de V1. (figure 3a, items en t1 groupés en fonction de V1, S1). Le lieu de l'occlusion est le plus antérieur pour les items en contexte /i/ ou /y/ ($p<.001$). Le lieu est légèrement plus postérieur en contexte /ɛ/ ($p<.01$). Les trois dernières voyelles entraînent un lieu encore plus postérieur ($p<.001$), avec, dans le cas de S2, une tendance encore plus postérieure pour /ɔ/ ($p<.02$). (/a/, /ɔ/ et /u/ seront appelées ci-dessous voyelles " postérieures "). Dans ces contextes, le lieu de l'occlusion est postérieur au 8° rang, et ne peut donc pas être représenté par les résultats EPG.

En ce qui concerne la magnitude des contacts (la somme des contacts sur le palais), chaque voyelle est caractérisée par une quantité propre, à l'exception, pour le sujet S1, du profil en contexte /u/ qui ne se distingue pas de manière significative ni de /a/ ni de /ɔ/. Pour le sujet S2, les profils en contexte /i/ et /y/ ne se distinguent pas, mais le profil en contexte /ɔ/ se distingue de ceux en /a/ et /u/.

Si l'on observe les différences sur chaque rang, on voit que pour S1, le profil en contexte /i/ se distingue des autres dès le 2° rang, et est similaire à /y/ sur les rangs 7-8. Le profil en contexte /ɛ/ ne se distingue de /y/ que sur les rangs 5, 7-8, et se distingue des voyelles postérieures sur les rangs 5-8. Pour les trois voyelles postérieures, /a/ et /u/ ne se distinguent jamais, et /ɔ/ se distingue des deux autres uniquement sur le 7° rang, et se distingue de /a/ sur le 6° rang et de /u/ sur le 8° rang. Chez le sujet S2, la distinction /i/-/y/ (avec plus de contacts pour /y/) n'est significative que sur les rangs 1-3. Le profil en contexte /ɛ/ ne se distingue des autres que sur le 8° rang, et de /ɔ/ à partir du 3° rang, mais reste toujours similaire à ceux de /a/ et de /u/.

En fonction de V2. Aucune différence n'est significative en fonction de V2 (figure 3b, items en t1 groupés en fonction de V2, S1). Les seules différences significatives sont obtenues en groupant les items en fonction du type (antérieur ou postérieur) de V2. Les profils diffèrent alors sur les 5° et 7° rangs pour S1, et des rangs 3-6 pour S2, avec davantage de contact pour les voyelles antérieures.

3.4 Caractérisation de la fin d'occlusion (t2)

En fonction de V1. Seul le sujet S1 présente des différences significatives entre les profils groupés en fonction de V1 en t2 (figure 4a). Ces différences concernent le 8° rang, le profil en contexte /i/ se différenciant de ceux de /a/ et de /ɔ/ et de celui de /u/. Si l'on groupe les items d'après le type (antérieur/postérieur) de V1, les deux sujets présentent des différences significatives. Pour S1, les différences se situent au niveau du 3° rang et des rangs 5-8 ($p<.002$). Pour S2, elles se situent sur les rangs 2-5 ($p<.008$) et sur les rangs 7-8 ($p<.001$).

En ce qui concerne la magnitude des contacts, on obtient des différences significatives en groupant les items selon le type de V1 ($p=.006$ (S1) et $p=.03$ (S2)).

En fonction de V2. La figure 4b groupe les items (en t2, S1) en fonction de V2. Les profils ne diffèrent significativement pour le lieu d'occlusion que pour les contextes /u, ɔ/ par rapport aux contextes /i, y/, ainsi que pour le contexte /ɛ/ qui diffère du contexte /ɔ/ pour le sujet S1 et de /u, ɔ/ pour le sujet S2. Le profil en contexte /a/ diffère aussi du contexte /ɔ/ pour le sujet S2.

Au niveau de la magnitude des contacts, les profils en contexte /i/ et /y/ ne diffèrent pas significativement, de même que les profils pour les trois voyelles postérieures. Le profil en contexte /ɛ/ ne diffère pas de /y/, ni de /u/ pour le sujet S1 et de /i/ pour le sujet S2.

Si l'on observe les profils rang par rang, on voit pour S1 que le profil en contexte /i/ ne diffère de celui de /y/ que sur le 2° rang, et de celui de /ɛ/ sur les rangs 2-4 et 7. Le profil en contexte /ɛ/ est similaire à ceux des voyelles postérieures sur les rangs 2-4, se rapproche des profils des voyelles antérieures sur le 5° rang, pour ensuite n'être différent que du profil en contexte /ɔ/ sur les rangs 6 et 8. Les profils en contexte des trois voyelles postérieures ne sont pas différents. Pour le sujet S2, les profils en contexte /i/ et /y/ ne diffèrent pas. Le profil en contexte /ɛ/ est similaire à celui de /i/ sur les rangs 2, 3 et 5-8. Sur les rangs 7-8, il ne se distingue que de /ɔ/ et /u/. Pour les profils en contexte de voyelles postérieures, seuls les contextes /a/ et /ɔ/ diffèrent sur le 8° rang.

4. DISCUSSION

4.1 Occlusives sourdes et sonores

Les résultats présentés en figure 1 (cf. 3.1) semblent montrer que l'occlusion est légèrement plus antérieure dans le cas de /g/. Cependant, il est également possible que l'étendue des contacts soit en fait plus importante pour l'occlusive sonore, ce que les données EPG ne nous permettent pas de constater puisqu'elles ne nous renseignent pas sur l'entièreté du contact langue/palais. Si ces résultats sont en opposition avec ceux de [Moo95], ils sont en accord avec ceux de [Dag94]. Ceci pourrait être dû au fait qu'en français, où le voisement des occlusives est généralement mieux soutenu qu'en anglais, on aurait un élargissement actif de la cavité buccale (cf. [Moo95] p.16).

4.2 Temporalité du geste d'occlusion

Le contact langue/palais apparaît comme plus important en t2 qu'en t1 sur les rangs postérieurs. Ces résultats semblent montrer une antériorisation du contact au cours de la production de l'occlusive vélaire. Notons toutefois que t1 est temporellement plus proche de V1 que t2 ne l'est de V2. En outre, nous ne pouvons faire l'hypothèse d'une symétrie entre le début et la fin du geste d'occlusion. Ces résultats témoignent en effet de l'avancement de la langue au cours de la production de ces consonnes, montré depuis [Hou67].

Cet avancement se produit dans le cas des items " symétriques " (qui ont des voyelles de même type de part et d'autre de la consonne, e.g. /akɔ/ ou /yki/). Par contre, pour les items " non-symétriques " (e.g. /ika/ ou /uky/), l'instant de l'occlusion qui a le contact le plus important sera celui qui est le plus proche temporellement

de la voyelle antérieure, c'est-à-dire t1 lorsque la voyelle antérieure est en V1, et t2 lorsque la voyelle antérieure est en V2. Ce résultat est similaire à celui de [Moo95], qui montrent que l'avancement de la langue n'est pas présent lorsque [i] est en V1, et est amplifié lorsque [i] est en V2 Ces constatations nous conduisent à considérer les interactions entre la temporalité du geste consonantique lui-même et l'influence au cours du temps du contexte vocalique.

4.3 Coarticulation

Si l'on compare l'influence qu'ont les voyelles sur l'occlusion à chacun des deux instants t1 et t2, on remarque qu'il y a moins de différences significatives en t2 selon V2 qu'en t1 selon V1. Au niveau de la magnitude des contacts, les profils en contexte /i/ et /y/, /y/ et /ɛ/, et /a/ et /ɔ/ diffèrent deux à deux en t1 selon V1, mais pas en t2 selon V2. On a également plus de différences significatives rang par rang entre deux profils en t1 selon V1 qu'en t2 selon V2, particulièrement au niveau des rangs postérieurs.

Ces résultats permettent de caractériser l'évolution dans le temps d'une occlusion soumise à l'influence des voyelles qui l'entourent. Il est important de constater qu'en t1, les profils du début de l'occlusion reflètent de manière significative la spécificité de chacune des 6 V1. On a donc affaire à un " continuum " de coarticulation selon V1, le début du geste d'occlusion s'adaptant facilement à la configuration de la langue en V1. Par contre, au relâchement de l'occlusion (t2), les profils présentent moins de configurations linguales différentes selon V2, et seules les voyelles les plus antérieures et les plus postérieures entraînent des distinctions significatives au niveau du lieu de l'occlusion. Ces résultats permettent de mieux comprendre la dynamique du geste d'occlusion, qui montre une coarticulation très forte en son début, mais qui, par l'évolution vers l'avant du contact langue/palais au cours de la production, est moins sujette à refléter des fines différences à son relâchement.

Le contact langue/palais pendant l'occlusion apparaît donc comme ayant une dynamique propre, dont l'interaction avec la coarticulation vocalique est assez complexe et varie au cours de la production. En effet, l'instant t2 montre davantage de différences selon V1 que l'instant t1 n'en montre selon V2 (cf. figures 3b et 4a). Ces interactions peuvent être précisées en testant l'égalité de variance de la magnitude de contact, en t1, entre deux groupes d'items qui diffèrent par le type (antérieur ou postérieur) de V1. Ces tests F montrent qu'en t1, on a des différences significatives selon que V1 est antérieure ou postérieure pour un sujet (p=.003 (S1), p=.054 (S2)). En t2, on n'a pas de différences significatives selon le type de V2 (p=.98 (S1), p=.79 (S2)). Ces résultats sembleraient montrer qu'en t1, les items qui ont une V1 postérieure varient plus que les items qui ont une V1 antérieure. Cette plus grande variance semble due au fait que les items en contexte de V1 postérieure vont varier en fonction du type (antérieur ou postérieur) de V2. Par exemple, une occlusive précédée de /a/ aura dès t1 un profil différent selon qu'elle est suivie de /a/ ou de /i/, avec davantage de contacts latéraux dans le second cas. En contexte de V1 antérieure, par contre, une occlusive précédée de /i/ ne sera pas différente en t1 selon qu'elle est suivie de /a/ ou de /i/.

Le contact langue/palais évolue vers l'avant au cours de la production. En comparant les figures 3a et 4b, on voit que cela se remarque davantage dans le contexte des voyelles postérieures que dans le contexte des voyelles antérieures, qui, par le contact plus important qu'elles supposent, entraînent un mouvement de moins grande amplitude. Ces résultats sont en accord avec les études antérieures (e.g. [Hou67] et [Moo95]), et placent donc dans une perspective de coarticulation le phénomène d'avancement de la langue comme dépendant du contexte vocalique. Cet avancement plus important en contexte de voyelle postérieure est également lié au phénomène de coarticulation par le fait qu'il détermine en partie l'organisation temporelle des gestes consonantiques et vocaliques en contribuant à réduire les distinctions que l'on peut faire d'après V2 au relâchement de l'occlusion.

CONCLUSION

En cherchant à décrire la palatalisation des occlusives vélaires du français par EPG, cette étude a caractérisé certains paramètres de l'organisation temporelle des phénomènes de coarticulation rencontrés. Les interactions entre différents gestes adjacents varient au cours de l'occlusion, et témoignent donc de la dynamique du contact langue/palais. Ces interactions ont précisé la nature du phénomène de palatalisation en fonction de la voyelle qui précède et qui suit, qui semble refléter la présence d'un continuum de coarticulation au début de l'occlusion, mais qui indique plutôt une différenciation entre les voyelles palatales et les plus postérieures au relâchement de l'occlusion.

Cette recherche est subventionnée par la Convention ARC " Dynamique des systèmes phonologiques ", 98-02 n° 226.

BIBLIOGRAPHIE

[Dag94] Dagenais P., Lorendo, L. et McCutcheon M. (1994). "A study of voicing and context effects upon consonant linguapalatal contact patterns", Journal of Phonetics, 22, 225-238.

[Har89] Hardcastle W.J., Jones W., Knight C., Trudgeon A. et Calder G. (1989). "New developments in EPG: a state of the art report", Clinical Linguistics and Phonetics, 1, 1-38.

[Hou67] Houde, R.A. (1967). A study of tongue body motion during selected speech sounds, PhD dissertation, University of Michigan

[Rou01] Rousselot P.J. (1901). Principes de phonétique expérimentale, Welter, Paris.

[Moo95] Mooshammer C., Hoole P. et Kühnert B. (1995). "On loops", Journal of Phonetics, 23, 3-21.

[Tes90] Teston B. et Galindo. B. (1990) "Design and development of a workstation for speech production analysis", Proceedings of VERBA 90: International conference on speech technology, Rome, 400-408.

Pression sous-glottique et débit d'air buccal des voyelles en français

Fabrizio Bucella°, Sergio Hassid+, Renaud Beeckmans°, Alain Soquet et Didier Demolin**

°Institut des Langues Vivantes et de Phonétique
*Laboratoire de Phonologie
+Hôpital Erasme
Université Libre de Bruxelles
Tél.: +32 2 650 20 18 - Fax: +32 2 650 20 07
E-mail: Fabrizio.Bucella@ulb.ac.be

ABSTRACT

According to classical view, the respiratory apparatus is a system which permits voluntary changing in intensity and maybe in the shape of the signal, but not (at least in the most known languages) as a system capable of raising the pressure for some given sounds. This paper examine variations going with different vowels and conclude that respiratory effort may be used to distinguish different sounds. Some French speakers seem to voluntary increase sub glottal pressure for some particular segments. Two subjects, one male and one female, participated to this study. We analysed different vowels productions for different controlled frequencies and different controlled intensities. We took measures of mean oral air flow *(dm^3/s)* and mean sub-glottal pressure *(hPa)*. Results showed (i) that sub-glottal pressure is lower for [a] than for [u] and for [i] ; (ii) that a vowel effect, as well for the pressure as for the oral air flow, was showed by repeated measures of analysis of variance. Obviously, the effect of the pressure reveals more complex mechanisms in the vowels production. In a first approach, this effect may be explained by three ways : (i) a difference in the air flow due to different vocal shapes ; (ii) a different respiratory control ; or (iii) a difference in the tension of the vocal folds. This last explanation seems to be an interesting track to explore for future experiments.

1. INTRODUCTION

Le système respiratoire est généralement considéré comme un système permettant de produire des variations volontaires de l'intensité et peut-être de la forme du signal, mais pas (au moins dans les langues les plus connues) comme un système capable de produire des augmentations de pression pour des sons particuliers [Tit94]. Tous les changements liés à des segments individuels, comme la chute de pression après un [h] ou l'augmentation de pression accompagnant l'occlusion du [k] sont considérés comme des aspects aérodynamiques du conduit vocal, sans contrôle volontaire. Ces différences peuvent être attribuées à des variations de la résistance au passage de l'air au travers des cordes vocales en vibration (l'impédance glottique) ou à la variation de la rigidité des parois du conduit vocal. Löfqvist [Lof75], résumant les observations faites sur l'activité respiratoire pour différentes catégories d'occlusives, conclut que, à l'exception des occlusives [fortis] en Coréen, les variations de la pression sous-glottique peuvent en général être attribuées à des variations de l'impédance glottique. Ce papier examine des variations accompagnant différentes voyelles et conclut qu'il n'y a pas que les Coréens qui utilisent l'effort respiratoire pour distinguer différents sons. Des locuteurs du Français semblent utiliser un accroissement volontaire de la pression sous-glottique pour des segments particuliers.

La relation entre pression sous-glottique et intensité des voyelles tenues du français a déjà fait l'objet d'une première analyse [Lec98a, Lec98b]. Une discussion plus détaillée sur la différence entre intensité produite et intensité perçue est donnée par la théorie des intensités spécifiques des voyelles [Ros71, Ros81 et Mar79] : des voyelles prononcées avec un effort constant et jugées comme isotoniques ne présentent pas la même intensité objective.

2. MATERIEL ET METHODE

Deux sujets ont participé à cette étude. Un sujet francophone masculin de 43 ans, un sujet francophone féminin de 29 ans. Les enregistrements ont lieu à l'unité ORL de l'hôpital Erasme, Université de Bruxelles. Les deux sujets ne présentaient aucune pathologie du larynx et n'étaient pas entraînés pour la tâche. La pression sous-glottique *(psg)*, la pression intra-orale *(pio)*, le débit d'air nasal *(dan)*, le débit d'air buccal *(dab)* et le signal de parole ont été enregistrés simultanément au moyen de la procédure suivante. Un tube flexible en caoutchouc (2 mm de diamètre interne) est introduit à travers la cavité nasale jusqu'à l'oropharynx pour la mesure de pression intra-orale. Une aiguille (2 mm de diamètre interne) est introduite dans la trachée après anesthésie locale (2% Xylocaine). Un tube de même type que le tube utilisé pour la pression intra-orale est connecté à l'aiguille. Le débit d'air buccal est mesuré avec un masque flexible en silicone. Le débit d'air nasal est mesuré au moyen d'une olive en silicone introduite dans la narine et reliée à un tube en plastique de 0.5 cm de diamètre. Tous les tubes et le masques sont reliés au

polyphonomètre, station d'acquisition automatique de paramètres aérodynamiques de la parole, développée à l'institut de phonétique d'Aix (France) [Tes90]. Cette station possède un module de post-traitement des données qui nous a permis de calculer l'intensité (RMS) du signal de parole et la fréquence fondamentale (méthode du peigne).

Le corpus se compose de mesures synchronisées de débit d'air buccal et de pression sous-glottique de voyelles. Les voyelles [a], [i] et [u] ont été tenues par le locuteur masculin et le locuteur féminin. Chacune de ces productions a été réalisée à 3 fréquences fondamentales et à 3 intensités différentes. Ce qui nous fait donc 9 types de répétitions de voyelles par locuteur.

Chaque voyelle a été segmentée en commençant au maximum de pression sous-glottique atteint après la mise en régime, correspondant au signal en régime stationnaire, et en terminant juste avant la chute de pression, correspondant à la fin du signal de parole (voir figures 1 et 2).

Sur ces portions de signaux nous avons pris les mesures du débit d'air buccal moyen et de la pression sous-glottique moyenne. Cependant les mesures de débit d'air buccal du locuteur féminin sont anormalement faibles, nous ne les avons pas retenues en raison d'un artefact possible de la mesure.

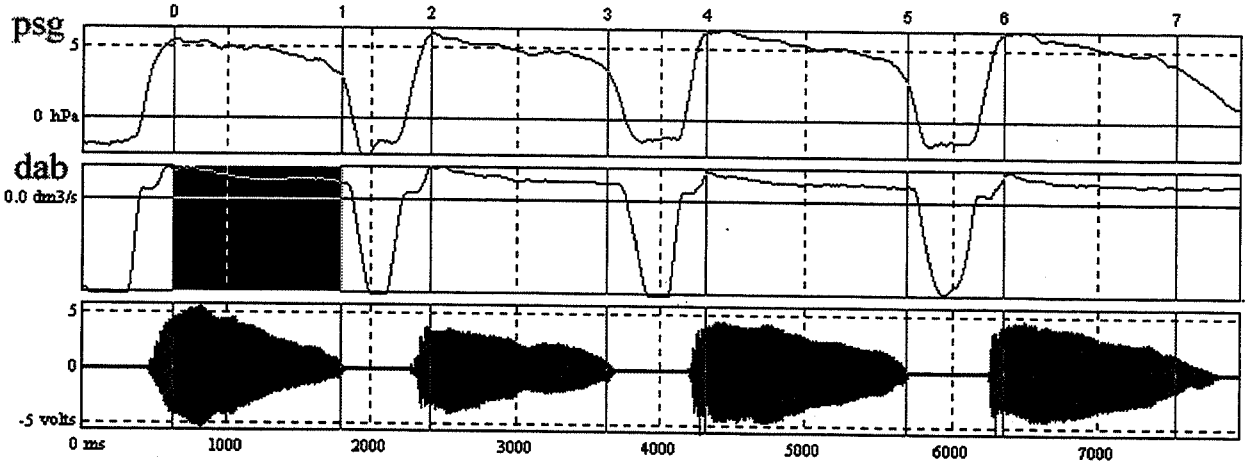


Figure 1: Pression sous-glottique (psg en *hPa*), débit d'air buccal (dab en dm^3/s) et représentation oscillographique du signal de parole d'un exemple de la voyelle [a] prononcée par le locuteur masculin.

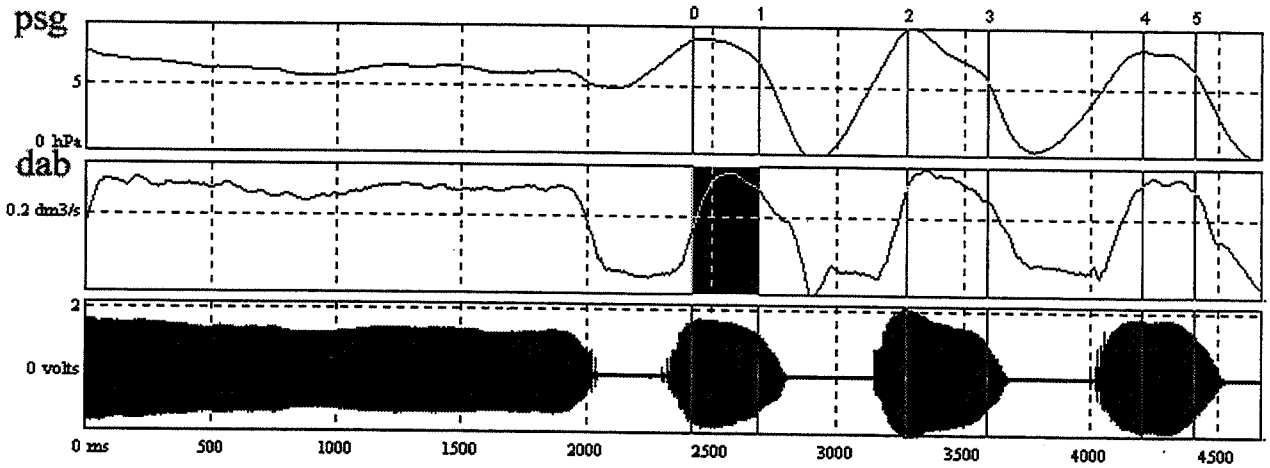


Figure 2: Pression sous-glottique (psg en *hPa*), débit d'air buccal (dab en dm^3/s) et représentation oscillographique du signal de parole d'un exemple de la voyelle [i] prononcée par le locuteur féminin.

3. RESULTATS

Sur les figures 1 et 2 sont représentés le signal de parole, le débit d'air buccal, (dm^3/s) et la pression sous-glottique (*hPa*) pour une production donnée. La figure 1 correspond à la production du [a] pour la fréquence la plus basse et l'intensité la plus faible du locuteur masculin. Le figure 2 représente la production du [i] à la fréquence la plus basse et à l'intensité la plus forte du locuteur féminin. Sur ces figures se trouve en impression plus foncée la portion segmentée du signal de débit d'air buccal, comme décrit précédemment.

Description statistique

Pour le locuteur masculin, les mesures prises sont le débit d'air moyen et la pression sous-glottique moyenne pour chaque répétition de voyelle dans les différentes conditions de production. Une mesure de la moyenne et de l'écart type sur l'ensemble des voyelles est aussi donnée. Enfin sont résumées et représentées sur le graphique les grandeurs statistiques moyenne et écart type global pour les voyelles [a], [i] et [u] - voir à ce propos les figures 3 et 4.

Pour le locuteur féminin, les mesures qui ont été gardées sont la pression sous-glottique moyenne pour chaque répétition de voyelle dans les différentes conditions de production. Une mesure de la moyenne et de l'écart type sur l'ensemble des répétitions de la voyelles est aussi donnée. Enfin sont résumées et représentées à la figure les grandeurs statistiques moyenne et écart type global pour les voyelles [a], [i] et [u] - voir à ce propos la figure 5.

Débit d'air buccal

Tableau 1: Résultats de l'analyse de variance de mesures répétées du débit d'air buccal pour le locuteur masculin. Comme le montrent bien les valeurs du p seul le facteur voyelle est déterminant.

Source	p **(signification < .05)**
Voyelle	.0001
Intensité	.6256
F0	.1233

L'ensemble du corpus de production du locuteur masculin a été soumis à une analyse de variance de mesures répétées (tableau 1) [Moo99, Ash93]. Les facteurs inter sujets de l'analyse de variance étaient les suivants : voyelle (trois possibilités : [a], [i] et [u]), fréquence (trois possibilités : basse, moyenne et haute) et amplitude (trois possibilités : faible, moyenne et forte). Un facteur intra sujet à quatre niveaux a aussi été introduit pour l'analyse répétée et correspondait à la mesure du débit d'air buccal émis sur une production. Le facteur voyelle s'est révélé le seul significatif valeur de $p = .0001 < .05$, seuil de signification.

Comme énoncé précédemment, une telle analyse n'a pas pu être réalisée sur le locuteur féminin vu les valeurs anormalement faibles des mesures.

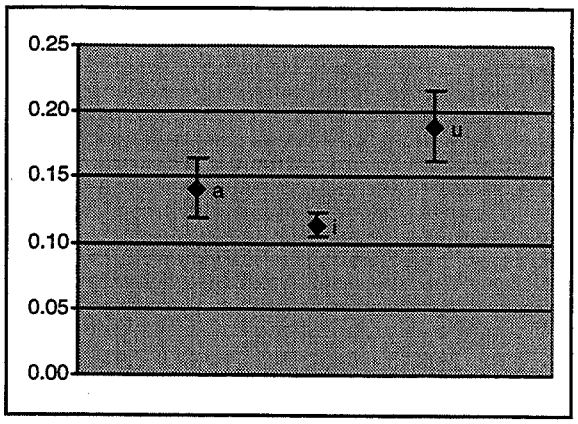


Figure 3: valeurs moyennes et barres d'erreur du débit d'air buccal (dab en dm^3/s) des voyelles [a], [i] et [u] calculées sur les trois ou quatre répétitions pour chacune des trois intensités et des trois fréquences (locuteur masculin).

Pression sous-glottique

Tableau 2: moyennes de la pression sous-glottique des locuteurs féminin et masculin. Lors de la production du [a] la pression est inférieure que lors de la production du [u] et du [i].

psg en [hPa]	[a]	[i]	[u]
Locuteur F	5.48	8.91	8.17
Locuteur M	5.74	7.45	7.22

Comme le montre le tableau 2, la pression sous-glottique est plus basse pour le [a] que pour le [u] et le [i]. Ceci se remarque pour le locuteur féminin et pour le locuteur masculin. Une analyse de variance de mesures répétées des locuteurs féminin et masculin a été réalisée. Les facteurs inter sujets de l'analyse de variance étaient les suivants : voyelle (trois possibilités : [a], [i] et [u]), fréquence (trois possibilités : basse, moyenne et haute) et amplitude (trois possibilités : faible, moyenne et forte). Un facteur intra sujet à quatre niveaux a aussi été introduit pour l'analyse répétée et correspondait à la mesure de la moyenne de la pression sous-glottique pour une répétition. Le facteur voyelle s'est révélé significatif valeur de $p = .000 < .05$, seuil de signification. Les facteurs fréquence et amplitude étaient aussi significatifs avec des valeurs de p semblables.

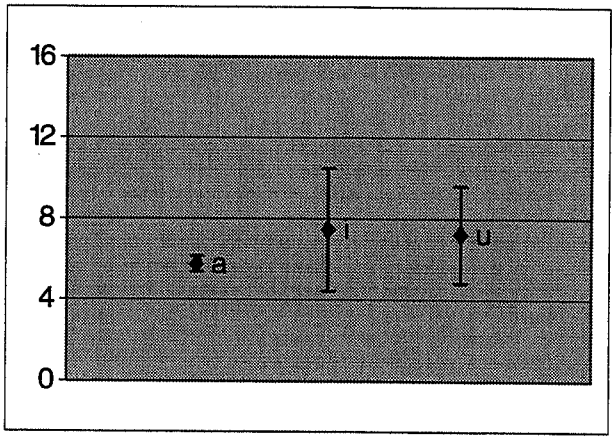


Figure 4: valeurs moyennes et barres d'erreur de la pression sous-glottique (psg en *hPa*) pour les voyelles [a], [i] et [u] calculées sur les trois ou quatre répétitions pour chacune des trois intensités et des trois fréquences du locuteur masculin.

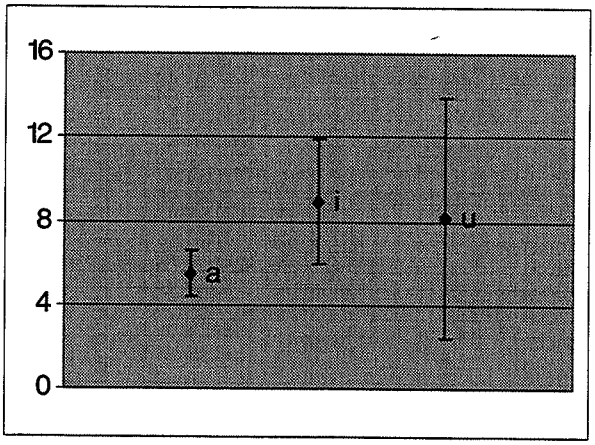


Figure 5: valeurs moyennes et barres d'erreur de la pression sous-glottique (psg en *hPa*) pour les voyelles [a], [i] et [u] calculées sur les trois ou quatre répétitions pour chacune des trois intensités et des trois fréquences du locuteur féminin.

DISCUSSION

Cette différence de pression sous-glottique entre [a], [i] et [u] est mise en évidence par l'analyse de variance de mesures répétées. Cet effet se remarque aussi pour la mesure du débit d'air buccal moyen. La différence de pression sous-glottique entre les voyelles est difficile a expliquer. Manifestement elle est révélatrice de mécanismes plus complexes dans la production des voyelles. En première approche une tel effet pourrait s'expliquer de trois manières : (i) une différence d'écoulement due à une forme différente du conduit ; (ii) un contrôle respiratoire différent ; (iii) une différence de tension dans les cordes vocales. Ce dernier effet semble être une des pistes à explorer pour le futur. La première hypothèse (i) ne permet pas de rendre compte de la différence du débit d'air buccal. En effet, en première approximation, sans couplage avec les cavités nasales, celui-ci devrait être sensiblement le même quelle que soit la forme du conduit. La deuxième hypothèse (ii) peut sans doute expliquer une partie du phénomène et devrait être explorée plus en détail. Par contre sachant que le cartilage thyroïde fait une rotation vers l'avant lors de la production du [i] et du [u], une différence de tension se produit sur les cordes vocales qui permet de rendre compte de la différence de pression sous-glottique moyenne mesurée - hypothèse (iii).

Cette recherche a été subventionnée par la convention ARC « Dynamique des système phonologiques » 98-02, n°226 de la Communauté Wallonie Bruxelles.

BIBLIOGRAPHIE

[Ash93] Ash, C., (1993) "The probability tutoring book: an intuitive course for engineers and scientists (and everyone else!)", IEEE, New-York, 470 p.

[Lec98a] Lecuit, V., Demolin, D., (1998) "The relationship between intensity and subglottal pressure with controlled pitch", Proceedings of the 5th I.C.S.L.P., Sydney, pp. 3079-3082

[Lec98b] Lecuit, V., Demolin, D., (1998) "Relation entre pression sous-glottique et intensité: étude des voyelles du français", Actes des XXIèmes J.E.P., Martigny, Suisse. pp. 299-302

[Lof75] Löfqvist, A., (1975) "A study of subglottal pressure during the production of Swedish stops", Journal of Phonetics, 3, pp. 175-189.

[Mar79] Marchal, A., Carton, F., (1979) "La pression sous-glottique: mesure et relation avec l'intensité et la fréquence fondamentale", in Semaine Larynx & parole, Institut de Phonétique de Grenoble – 8-9 fév. 1979, pp. 315-327.

[Moo99] Moore, D.S., McCabe, G.P., (1999) "Introduction to the practice of statistics", W.H. Freeman and Company, New-York, 825 p.

[Ros71] Rossi, M., (1971) "L'intensité spécifique de voyelles", Phonetica, 24, pp. 129-161.

[Ros81] Rossi, M., Autesserre, D., (1981) "Movements of the hyoid and the larynx and the intrinsic freqeuncy of vowels", Journal of Phonetics, 9, pp. 233-249.

[Tes90] Teston B., Galindo B. (1990) "Physiologia : un logiciel d'analyse des paramètres physiologiques de la parole", Travaux de l'Institut de Phonétique d'Aix, 13, pp.197-217.

[Tit94] Titze, I.R. (1994) « Principles of voice production », Prentice Hall, U.S.A., 354 p.